Grace Yang, gy654, N10610063

5.9.2022

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

1. Background: general information about your chosen ADS

A. What is the purpose of this ADS? What are its stated goals?

The ADS we propose to analyze in this project is a credit scoring algorithm whose stated goal is to make predictions about loan applicants' probability of default (i.e., the probability that someone will experience financial distress in the following two years). The ADS will assign either a positive outcome (SeriousDlqin2yrs= 0) or a negative outcome (SeriousDlqin2yrs = 1) to a person based on an array of input features manifesting the person's credit history and financial capabilities. The ADS should be transparent enough to help us determine typical features and patterns of behavior that lead to a future inability to make debt repayments. Banks can utilize this information to determine whether or not a loan should be granted. Borrowers can use it to help make the best financial decisions and increase their chances of obtaining a loan. While prediction accuracy should be prioritized, we should also make sure that the ADS will not deny someone a loan based on sensitive attributes such as age. The ADS should also be robust to linkage attacks such that the privacy of borrowers' financial data is under protection.

B. If the ADS has multiple goals, explain any trade-offs that these goals may introduce.

The ADS we choose to analyze has three goals and each goal may incur some trade-offs as follows:

1. predictive accuracy vs. fairness to subgroups

An overarching goal of the ADS is to increase the overall accuracy in predicting the probability that someone will experience financial distress in the next 2 years. The accuracy is defined as the proportion of true results among the total number of cases examined. ((TP + TN)/Total). When accuracy is optimized, the bank could identify more candidates that will truly experience financial distress in the next two years. Correctly identifying individuals with high risks of default may prevent banks from financial losses since the bank can get the loan back in time with fewer people defaulting. Correctly identifying individuals with low risks of default may increase the bank's profit since the bank can be more confident in granting them more loans being assured that they have the financial capability of repaying. The bank can generate more revenue from the interest it charges more applicants with. However, it should be noted that the false-negative rate in certain subgroups is likely to increase as we optimize accuracy. If this is the case, the new identification process will unfairly mark someone as having a high default risk due to their young age for instance. Thus, the ADS model will decrease the number of loans lent out to borrowers from certain subgroups. This may incur several trade-offs in which the bank will become warier in lending and experience a decrease in the profit from the loans, and individuals may not receive the loan despite a good financial standing. Similarly, if the false positive rate in certain subgroups is likely to increase as we optimize accuracy, then the ADS will unfairly mark someone from this group as having low default risk despite their worrying financial state. Then the bank's profit would decrease due to these individuals' unanticipated default in the future.

2. Sensitive data protection vs. accuracy

The ADS should protect loan applicants' sensitive data and ensure that their privacy is unviolated. Stipulated by civil law, some sensitive attributes such as applicants' names, SSN, gender, race, and zip codes are masked for this purpose. Differential privacy should be achieved such that no information about a specific individual can be learned by querying the dataset. We cannot identify an individual through linkage attacks, which is exposing an individual by joining the dataset with auxiliary datasets with identifying properties. The public exposure of an applicant's financial data harms the individual involved since they are likely to be denied other opportunities if this information is made public, or they are likely to be discriminated or humiliated against. These victims may be also vulnerable to annoying advertisement delivery with misleading financial solutions. While privacy protection is important, optimizing privacy by masking sensitive attributes or injecting perturbations may reduce the overall accuracy of the ADS. Some attributes such as gender or race may contain information about an individual's likelihood of financial distress.

3. Transparency vs. its potentiality for exploitation

Rather than a black box generating perplexing outcomes, the ADS should be transparent enough to help us determine typical features and patterns of behavior that lead to a future inability to make debt repayments. Banks can utilize this information to determine whether or not a loan should be granted. They can provide an individual with a reasonable explanation of their decision of granting or denying his or her loans. Borrowers can use it to help make the best financial decisions and increase their chances of obtaining a loan.

A potential tradeoff is that an individual can exploit this transparency in information and deceive a bank for a loan by fabricating his financial profile according to the interpretation provided by the ADS.

4. Fitting to the static data vs. the dynamic nature of people and economy

While fitting data is important, a trade-off of accuracy in predicting future default rates would occur if the historical data is too outdated for trend prediction. The scoring system will evaluate future behaviors based on the past patterns extracted from 250,000 borrowers. The static nature of the rating is done by utilizing historical data of the borrowers at a specific duration of time. Factors such as policy, the robustness of the economy, social stability, and environmental status have a certain degree of direct and indirect effect or influence on the borrowers. Thus, any shifts after the construct of the model to the end of providing more accurate credit scores will defeat the very goal of the rating.

5. Predicting 2-year financial distress vs. predicting long-term financial distress

The prediction of the model is for the future 2 years. However, a loan usually lasts longer than 2 years. Thus, the usefulness of the model could be questioned. Rating candidates' credit by taking account of the next two years' financial statuses is not enough to account for one's possibility to default on a loan that has a period longer than 2 years. For example, if one is viewed as a valid candidate to receive a loan that lasts 10 years, and the candidate goes defaults on the 5th year, then the bank fails to minimize the default risk using the prediction of this ADS. This presents the trade-off between optimizing predictive accuracy in 2-year default rates and long-term creditworthiness.

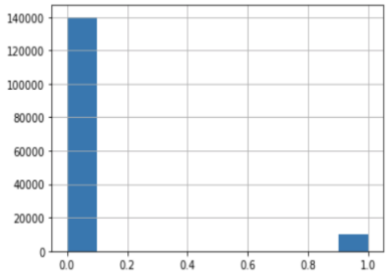
Section2. Input and output

A. Describe the data used by this ADS. How was this data collected or selected?

The training dataset used by this ADS contains 150,000 entries. Each entry has corresponding values for 10 features (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). Each entry in the training dataset is also attached to a label indicating if the person experienced 2-years of serious delinquency.

The test dataset contains 101503 entries. Each entry has values for the aforementioned 10 features, but the binary label SeriousDlqin2yrs for each data remains to be determined. The goal of this ADS is to make predictions on the label SeriousDlqin2yrs using a fair classifier. The Kaggle competition webpage where data can be downloaded provides no information about how the data is collected or selected. The lack of context is problematic because underlying biases can be undetected if the source of the data remains untransparent. Specifically, we do not know the participating population from which the credit data is drawn. Applicants' financial situation may vary depending on countries and socioeconomic classes. Without metadata informing us about participants' general information, the ADS might be biased in reproducing fair results across subgroups since some groups may be less represented in the dataset. We also do not know the freshness of the data, so we cannot be certain that the ADS model built based on the training data would generate timely credit prediction results and can be deployed in the real world. We also have little knowledge about the data collection techniques which can often influence the data. If the sampling methodology is inappropriate in capturing the applicant's data, then biases are introduced during the sampling stage, then the consistency of our data is reduced, thus undermining the conclusion we draw based on these data. We do not know how data is entered manually into the dataset. If some outliers in RevolvingUtilizationOfUnsecuredLines and DebtRatio might be due to registering errors, then it is justifiable to drop them. Otherwise, dropping outliers may have the effect of reducing accuracy since some useful information on credit default is left unexploited.

B. For each input feature, describe its datatype and give information on missing values and the value distribution. Show pairwise correlations between features if appropriate. Run any other reasonable profiling of the input that you find interesting and appropriate.



Label: SeriousDlqin2yrs (Y/N)

Definition: Person experienced 90 days past due delinquency or worse

Overall data description:

0 139974
1 10026

Note that the base rate of delinquency after pre-processing is 0.07169803012065878, which means a classifier that assigns 0 to each entry will achieve a 93% accuracy, so the accuracy score is not sufficient to evaluate the model's performance, we would also like to make the model more transparent by inspecting how each factor contributes to the prediction.

Input feature1: RevolvingUtilizationOfUnsecuredLines (percentage)

- Definition: The total balance on credit cards and personal lines of credit except for real estate and no installment debt like car loans divided by the sum of credit limits.
- Overall data description:

count	mean	standard deviation	min	25%	50%	75%	max
150000	6.048438	249.755371	0.000000	20.029867	0.154181	0.559046	50708

There is no missing value in this feature.

We follow the exploration already done in the ADS notebook. Examine RULL values close to 1 and observe at what rate they are defaulting. We start with 0.9 to 4.0. These 20,000 people are defaulting at a rate of almost 1 in 4.

count	mean	standard deviation	min	25%	50%	75%	max
19805.000000	0.225347	0.417821	0.000000	0.000000	0.000000	0.000000	1.000000

There are only 23 records with RUUL between 4 and 10, but they're still defaulting at a high rate.

count	mean	standard deviation	min	25%	50%	75%	max
23.000000	0.260870	0.448978	0.000000	0.000000	0.000000	0.500000	1.000000

Consider those with RUUL > 10.

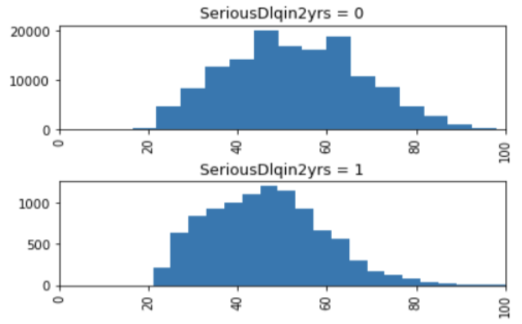
count	mean	standard deviation	min	25%	50%	75%	max
241.000000	0.070539	0.256587	0.000000	0.000000	0.000000	0.000000	1.000000

These 241 people are not defaulting any more than anyone else, despite some of them owing hundreds of thousands of times their credit limits. These seem to be inconsistent with the rest of the data, so we can remove them from our model. Dropping data with RevolvingUtilizationOfUnsecuredLines >10 seems valid since it does not arbitrarily change demographic group proportion and leaves out people with higher default rates.

Input feature2: age(integer)

- Definition: Age of borrower in years
- Overall description:

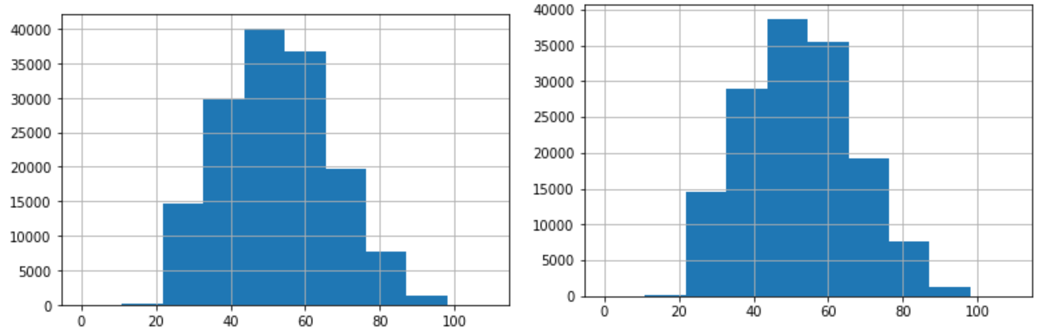
count	mean	standard deviation	min	25%	50%	75%	max
150000	52.295207	14.771866	0	41.000	52.000	63.000	109.000



There is no missing value in this feature.

The distribution of people that defaulted versus those who did not show us that, generally, younger people were more responsible for defaulting. We will examine the fairness of the ADS in risk assessment and see how different kinds of errors are distributed among sub-populations (age>=55 vs. age<55) later in this report. We need to confirm that in our data set old age has a negative influence on the probability of default. Then we can assume that the models we build reflect that trend and won't deny someone a loan simply because they are young.

The distribution of age of the cleaned dataset is similar to the distribution of age of the original dataset. Therefore, the data filtering process does not change the underlying distribution of age and makes either young people or old people more represented in the dataset.



The distribution of age in the original dataset (left)

The distribution of age in the pro-processed dataset (right)

Input feature3: Debt Ratio (percentage)

- Definition: Monthly debt payment, alimony, living costs divided by monthly gross income
- Overall description:

count	mean	standard deviation	min	25%	50%	75%	max
150000.000	353.005076	2037.818523	0.000000	0.175074	0.366508	0.868254	329664.00

There is no missing value in this feature. However, an examination of extreme values informs us that 2.5% of the dataset owes more than 3,500 times what they own. We investigate further to see if these are outliers or not. We see two particularly concerning things: The first is that among 4,000 records with DebtRatio > 3,500, only 185 of them have a value for monthly income. Further, the people who do have monthly income seem to either have a monthly income of either 1 or 0. We see that 164 of those 185 entries, 164 of them have the same value for 2-year default rate and monthly income, indicating that there is a data-entry error.

Despite owing thousands of times what they own, these people aren't defaulting any more than the general population. We can conclude that these entries must be data-entry errors, so we will remove them from our model. Debt_ratio outliers (people with DebtRatio > 3489.025) are first removed. Then data with nan DebtRatio value is filled with the median of DebtRatio after removing outliers. This seems valid since the median is a better metric to impute missing incomes than the mean given its insensitivity to extreme values.

Input feature4: MonthlyIncome(real)

- Definition: Monthly income
- Overall data description:

count	mean	standard deviation	min	25%	50%	75%	max
1.202690e+05	6.670221e+03	1.438467e+04	0.00e+00	3.4000e+03	5.4000e+03	8.2490e+03	3.00875e+06

There are 29731 missing values in this feature. However, of 4,000 records with DebtRatio > 3,500, only 185 of them have a value for monthly income. Further, the people who do have monthly income seem to either have a monthly income of either 1 or 0. We see that of those 185 entries, 164 of them have the same value for 2-year default rate and monthly income, indicating that there is a data-entry error. Despite owing thousands of times what they own, these people aren't defaulting any more than the general population. The ADS is justified to conclude that these entries must be data-entry errors, so it removes them from the model.

Input feature5: NumberOfOpenCreditLinesAndLoans (integer)

- Definition: Number of open loans (installment like car loan or mortgage) and lines of credit
- Overall data description:

count	mean	standard deviation	min	25%	50%	75%	max
150000.00	8.452760	5.145951	0.000000	5.000000	8.000000	11.000000	58.000000

There is no missing value in this feature.

Input feature6: NumberOfTime30-59DaysPastDueNotWorse (integer)

- Definition: Number of times borrower has been 30-59 days past due but no worse in the last 2 years
- Value counts:

0	1	2	3	4	5	98	6	7	8	9	96	10
126018	16033	4598	1754	747	342	264	140	54	25	12	5	4

12	13	11
2	1	1

There is no missing value in this feature. However, it is concerning that no one is between 14 and 96 times late. We will deal with those values later aligned with input feature 8. The ADS winsorizes by replacing all the 96/98s with 18 to make them not extreme outliers.

Input feature7: NumberOfTime60-89DaysPastDueNotWorse (integer)

- Definition: Number of times borrower has been 60-89 days past due but no worse in the last 2 years
- Value counts:

0	1	2	3	98	4	5	6	7	96	8	11	9
142396	5731	1118	318	264	105	34	16	9	5	2	1	1

There is no missing value in this feature. However, it is concerning that no one is between 12 and 96 times late. We will deal with those values later aligned with input feature 8. The ADS winsorizes by replacing all the 96/98s with18 to make them not extreme outliers.

Input feature8: NumberOfTimes90DaysLate(integer)

- Definition: Number of times borrower has been 90 days or more past due
- Value counts:

0	1	2	3	4	5	6	7	8	9	10	11	12
141662	5234	1555	667	291	131	80	38	21	19	8	5	2
13	14	15	17	96	98							
4	2	2	1	5	264							

Note that no one is between 17 and 96 times late, but hundreds of people are 98 times late. We can take a look at these few hundred records. Somehow, all these (300) people were 30-59 days late 96/98 times, 60-89 days late 96/98 times, and 90+ days late 96/98 times. This is concerning. However, the data might not be garbage, because (as expected) these people are defaulting at a massive rate (55%) compared to the population (6%). Therefore, we don't want to throw away this data.

The ADS winsorizes by replacing all the 96/98s with 18 to make them not extreme outliers and see if that improves the models. We don't expect that to improve the random forests, which are robust to outliers, but it might improve the SVMs.

Input feature9: NumberRealEstateLoansOrLines (integer)

- Definition: Number of mortgage and real estate loans including home equity lines of credit
- Overall data description:

count	mean	standard deviation	min	25%	50%	75%	max
150000.000000	1.01824	1.129771	0.000000	0.000000	1.000000	2.000000	54.000000

There is no missing value in this feature.

Input feature10: NumberOfDependents(integer)

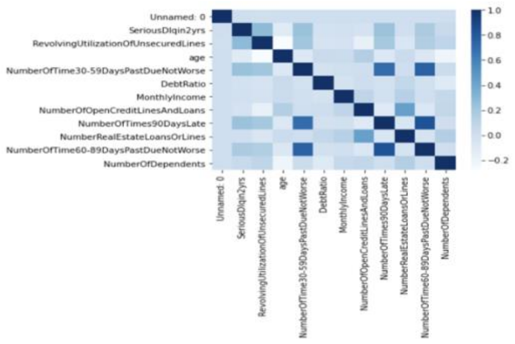
- Definition: Number of dependents in family excluding themselves (spouse, children, etc)
- Value counts:

0.0	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	10.0	13.0	20.0
86902	26316	19522	9483	2862	746	158	51	24	5	5	11	

There are 3924 missing values in this feature in the original dataset, but after data cleaning, there is no missing value in this feature.

Correlation:

Attached here is a correlation heatmap between 11 features. Note that some features such as “NumberOfTime60-80DaysPastDueNotWorse” and “NumberOfTime90DaysLate” are highly correlated.



C. What is the output of the system (e.g., is it a class label, a score, a probability, or some other type of output), and how do we interpret it?

The ADS produces the output of a binary label SeriousDlqin2yrs, which indicates whether a person is predicted to experience 90 days past due delinquency or worse based on the provided predictors. Each entry is assigned to one binary label as the model output. The binary label is represented by dummy variables 0 and 1. A 0 indicates that the person will not experience 90 days past due delinquency or worse and a 1 indicates that the person will experience 90 days past due delinquency or worse.

The notebook that we decide to use for our analysis does not provide an output file or generate binary labels of SeriousDlqin2yrs using the test sets for each entry. It only compared the performance of various models, including the random forest model and simple SVM model. Therefore, we split the training dataset into training and testing sets and add some code to run the model on the test sets. We make predictions on the test set using the model that proved in the ADS to have the best performance, which is the random forest model with parameters n_estimators=16, max_depth = 9, random_state=0, and the dataset that has removed the utilization outliers. We dropped the entries in the test set with null values.

After splitting, the training dataset has 102214 entries, and the test dataset has 43807 entries. The value counts of the output results are as follows:

```
outcomes:
0    42904
1      903
```

This result suggests that among the 43807 test entries, 42904 people receive a label of 0, meaning they will not experience 90 days past due delinquency or worse and 903 receive a label of 1, meaning they will experience 90 days past due delinquency or worse. Compared to the true label, we conclude that the model has an accuracy of 0.9377496747095213 and an AUC of 0.5865963364026058

Section3: Implementation and validation:

Present your understanding of the code that implements the ADS. This code was implemented by others (e.g., as part of the Kaggle competition), not by you as part of this assignment. Your goal here is to demonstrate that you understand the implementation at a high level.

A. Describe data cleaning and any other pre-processing

Based on all of these insights we gained from the data exploration of input features, the ADS proposes three data cleaning pre-processing steps that will reduce errors and potentially improve the model’s performance.

- The ADS can remove all outliers in the DebtRatio feature (values > 3489.025).
- The ADS can fill in nulls in DebtRatio with the median of the rest of the DebtRatio in the dataset.
- The ADS can Drop all the entries with outliers in the RevolvingUtilizationOfUnsecuredLines feature (values >10)
- The ADS can replace outliers in ‘NumberOfTime30-59DaysPastDueNotWorse’, ‘NumberOfTime60-89DaysPastDueNotWorse’, and ‘NumberOfTimes90DaysLate’ (values >90) with a reasonable guess of 18.

The ADS produces four datasets according to the different modes of modification mentioned above.

1. Median fill DebRatio with DebtRatio outliers removed (removed_debt_outliers)
2. Median fill DebtRatio with DebtRatio outliers removed and with NumberOfTime30-59/60-80/90Pass DueNotWorse outliers removed (dfn98)
3. Median fill FebRatiio with DebtRatio outliers removed and with RUUL outliers replaced. (dfus)

B. Give high-level information about the implementation of the system

The system first performs some basic data exploration steps to give readers a general idea about the distribution of the input data. It marks cases where data is dubious or concerning and needs to be preprocessed before being used as input to the model.

Then the system performs data cleaning procedures and produces four training sets according to different modes of modification for model selection.

The system then develops a tester program that automatically runs k-fold cross-validation on various scikit models across different datasets. The modified datasets have a big influence on the SVM AUC scores, but they do not affect the Random Forests as much. However, there is a noticeable gain in the performance between simply dropping the missing values and the modified datasets across all models. The AUC of the Simple Random Forest is higher than the AUC of Simple SVM, so the ADS decides to move forward with the random forest and figure out the best tuning parameters using the grid search. The best model is the random forest with a depth of 9 and 16 estimators. When we use the data set that removes the RULL outliers, it gives us an AUC of 0.8662. The ADS also tests a simple K-Nearest Neighbors model, but the AUC is relatively low (0.52).

c. How were the ADS validated? How do we know that it meets its stated goal(s)?

The ADS is validated by performing k-fold cross-validation on the training dataset. The ROC-AUC scores of different models and different datasets are then compared. The model on the dataset with the highest ROC-AUC validation score is considered the one with the best performance in meeting its stated goal of predicting whether an applicant would be able to make debt repayments in the future. However, the validation process of the ADS is hard to trace since it is defined and made automated by a systematic pipeline. To better analyze the output of the system, we separate the model and fit it on the suggested dfus dataset. We make predictions on a subset of the training dataset with labels removed.
A glimpse of the performance of the model:

```
Accuracy:  0.9377496747095213
AUC: 0.5865963364026058
```

Section 4. Outcome

A. Analyze the effectiveness (accuracy) of the ADS by comparing its performance across different subpopulations.

Based on the result of model testing, the best ADS selected is the random forest with a depth of 9 and 16 estimators. We will analyze the accuracy of the ADS by comparing its performance on young people vs old people (cutoff age = 55). We encode people with age<55 as 0 belonging to the unprivileged class and encode people with age>= 55 as 1 belonging to the privileged class.

- Since Kaggle does not provide a validation dataset, we can split the training dataset into one dataset for model training and one for model testing.

```
size of training dataset:  102214
size of test dataset:  43807
prediction outcomes:
0    42904
1      903
```

- We calculate two fairness metrics (mean_difference and disparate_impact) on the training data. We put our data back into an aif360 dataset format so that we can use all of the fairness metrics provided by the package. Calculate some fairness metrics for `orig_aif360` and `preds_aif360`: mean difference and disparate impact.

```
Results:
Train original metric: Mean difference = -0.052929
Train original metric: Disparate impact = 0.945005
Train predicted metric: Mean difference = -0.022389
Train predicted metric: Disparate impact = 0.977414
Test original metric: Mean difference = -0.051321
Test original metric: Disparate impact = 0.946730
Test predicted metric: Mean difference = -0.022865
Test predicted metric: Disparate impact = 0.976958
```

We choose disparate impact as a metric to evaluate the fairness of the model to different subgroups. It is the probability of individuals belonging to the unprivileged group receiving a positive outcome over the probability of individuals belonging to the privileged group receiving a positive outcome. The value smaller than one verifies the definition of privileged vs unprivileged groups. The model turns out to mitigate bias slightly. 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. The metric approaching 1 suggests that the output of the model is fairer in distributing positive outcomes between the privileged group and the unprivileged group.

We also choose the mean difference to evaluate the fairness of the model to different subgroups. It is a standard statistic that measures the absolute difference between the mean value of prediction in two groups. The negative value of the mean difference verifies the definition of privileged vs unprivileged groups. The model turns out to mitigate bias slightly. 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. The shrinking of the metric in absolute values suggests that the output of the model is fairer in distributing positive outcomes between the privileged group and the unprivileged group.

- Since we have true values and predicted values, we can compare the true positive rate and false positive rate by subgroups. We can use the “ClassificationMetric” function to do this.

Results:

Error rate difference (unprivileged error rate - privileged error rate) = -0.044058
False negative rate for privileged groups (age <55) = 0.009039
False negative rate for unprivileged groups (age >=55) = 0.001520
False negative rate ratio = 0.168131
False positive rate for privileged groups (age<55) = 0.746940
False positive rate for unprivileged groups (age>=55) = 0.806471
False positive rate ratio = 1.079700

We identified bias in the training data. We should therefore not find it surprising that we have a bias in a model trained on that data. The false-positive rates between the privileged groups and the unprivileged groups are similar (0.747 and 0.806). The two groups have a comparable probability of being marked as having a low risk of delinquency.

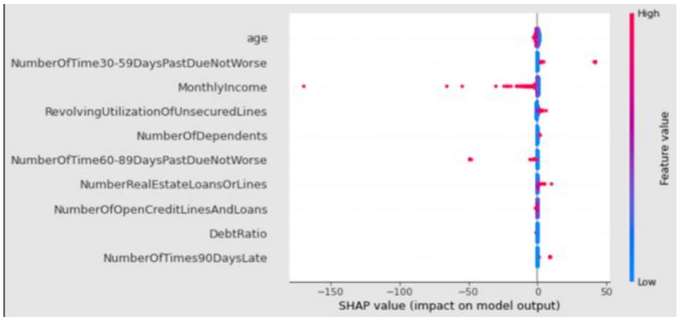
Though the false-negative rates in both groups are small. the false-negative rate for the unprivileged groups is 0.16 times the false-negative rate for the privileged groups. The model will be slightly more likely to overestimate the default risk of people with an age smaller than 55 than of people with an age greater than 55, but the model is overall fair across different age groups. The FPR and FNR are within an acceptable range.

B. Select one or several fairness or diversity measures, justify your choice of these measures for the ADS in question and quantify the fairness or diversity of this ADS.

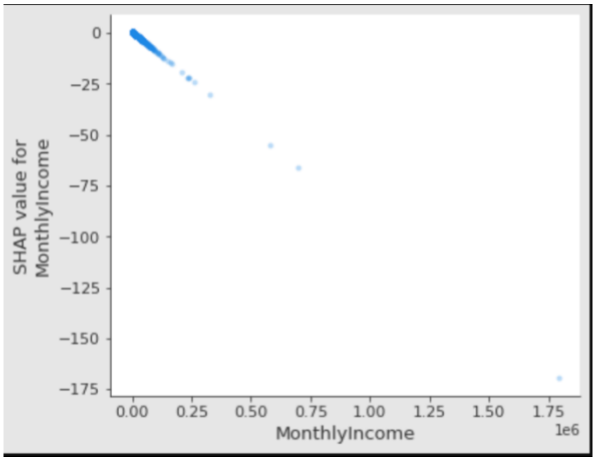
The input of the data needs to be fair and equally representative both privileged and unprivileged to allow ADS to produce a fair assessment result. We also need to check the fairness of the predicted data from the model and compare them with the fairness metrics we calculated before to see the fairness of the model. As mentioned in the above question, we transformed the input data and rendered the group with age above 55 as the privileged group and the group with age below 55 as the unprivileged group. We calculated several fairness measures on the privileged and unprivileged groups in the dataset of our ADS. We calculated the mean difference, which by definition, is comparing the percentage of favorable results for the privileged and unprivileged groups. The mean difference will provide us the information on whether the unprivileged are at the disadvantage of receiving a favorable outcome. A mean value equal to 0 will suggest that the privileged group and unprivileged group have an equal outcome. A mean value less than 0 will then suggest that the input data of our ADS is biased and will have disadvantages for the unprivileged group. We also calculate the disparate impact, which by definition is the ratio of predicted favorable outcomes for the unprivileged group compared to the privileged group. A disparate impact value of less than one will indicate less fairness as privileged groups have more favorable outcomes. As already calculated in the above question when comparing the performance of the model across different subpopulations, there exists a slight bias between the privileged group and unprivileged the input data as the mean difference of -0.05 and the disparate impact of 0.095 in the input data. The model mitigated such bias and made the model fairer by reducing the mean difference to -0.02 and disparate impact to 0.98.

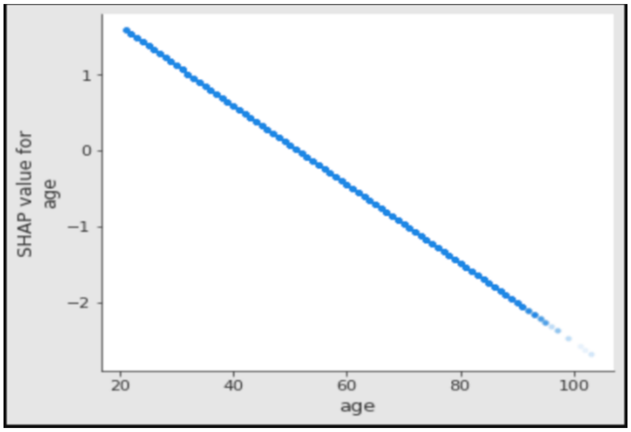
C. Develop additional methods for analyzing ADS performance: think about stability, robustness, performance on difficult or otherwise important examples (in the style of LIME), or any other property that you believe is important to check for this ADS.

We further analyze the transparency of ADS by checking the feature importance. We employ the SHAP package on our training data set. We fit the training set to a simple logistic regression model and construct a SHAP linear explainer from it. We plot the SHAP values on the illustrative summary plot which helps us identify features that may have a significant influence in determining whether a person will suffer economic distress, which is the binary label for the attribute SeriousDlqin2yrs.



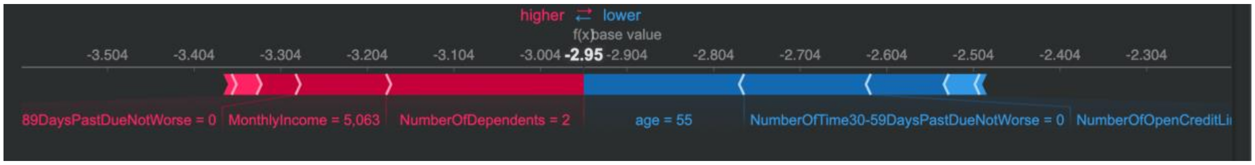
Moreover, we can check a single feature and observe its importance and relationship to the outcome vary across its range. Below are instances of “MonthlyIncome”.





In addition, we check whether the characteristics of young age will positively or negatively influence the outcome of the model. We generate a dependence plot of age to observe the relationship. A larger age corresponds to a more negative SHAP value, contributing negatively to the outcome towards 0, which is a label of lower risk of financial distress.

Moreover, for a specific entry, we check the direction of the contribution of each important feature to the outcome. For instance, for the third entry in the test dataset, the individual has a sum of -2.95, indicating that he or she has a low risk of financial distress in two years.



We can inspect this individual’s data for more details by analyzing its SHAP values. This procedure improves the interpretability of the model and enables the bank to provide the loan applicant with a reasonable explanation of the decision of either granting or denying the loan. The borrower can use this information to help make the best financial decisions and increase their chances of obtaining a loan. They know what features to work on to become a better loan candidate in the future.

```
Features contributing positively:
[('NumberOfDependents', 0.22813602362019433),
('MonthlyIncome', 0.10465104216082206),
('NumberOfTime60-89DaysPastDueNotWorse', 0.04501833142453436),
('DebtRatio', 0.028808927297183585),
('NumberRealEstateLoansOrLines', 0.009640522460943797)]

Features contributing negatively:
[('NumberOfTimes90DaysLate', -0.008422658057246143), ('RevolvingUtilizationOfUnsecuredLines', -0.035657661182161654),
('NumberOfOpenCreditLinesAndLoans', -0.09005234920450421),
('NumberOfTime30-59DaysPastDueNotWorse', -0.14633528218103484),
('age', -0.18519998073909322)]
```

Section5: Summary

A.Do you believe that the data was appropriate for this ADS?

It is reasonable to include many predictive features in the dataset since they are valuable in predicting the probability of delinquency. Specifically, the debt-to-income ratio is a good indication of an applicant's ability to repay the loan. The lower the DTI, the less risky a person is to banks since he has more disposable income to pay the debt. NumberOfTimes90DaysLate is a good metric of a person's credit repayment history, which is a good indication of the applicant's future repayment behaviors. A high score in the NumberOfDependents indicates that the person has more financial burdens, reducing his or her ability to repay the loan.

The dataset does not contain sensitive attributes including gender and race, reducing the likelihood that the ADS would produce disparate impacts on different subgroups concerning genders and races.

Some features are highly correlated (NumberOfTimes30-59DaysLate, NumberOfTimes60-89DaysLate), so we need to filter through redundant information to gain insights.

The dataset is anonymized. Personal identifiers including name, SSN, mailing address, and zip code are masked. However, if auxiliary information is available for analysts to pinpoint individuals in the dataset with a linkage attack, then the ADS cannot protect an individual's privacy since it is not robust to reconstruction attacks.

B. Do you believe the implementation is robust, accurate, and fair? Discuss your choice of accuracy and fairness measures and explain which stakeholders may find these measures appropriate.

We measure the accuracy score of the ADS (TP +TN)/total = 0.9377496747095213. Note that the base rate of delinquency after pre-processing is 0.07169803012065878, which means a classifier that assigns 0 to each entry will achieve a 93% accuracy, so the accuracy score is not sufficient to evaluate the model's performance. We analyze its accuracy on people from different age groups and conclude that the model is robust and distributes the positive outcome relatively fairly based on similar FPR and FPR across subgroups, mean difference approaching 0, and disparate impact ratio approaching 1. In terms of the robustness of the model, unfortunately, we do not have sufficient information on whether there are auxiliary datasets that may incur the risk of linkage attack. If such a dataset exists, then the ADS needs improvement in robustness such as using randomized response and differential privacy, otherwise, the system is robust in protecting data contributors' privacy. We also need a bigger dataset to confirm that the model would generate consistent predictions. We increase the robustness of the model and render it more transparent by inspecting how each factor contributes to the prediction.

Banks may find the optimization of accuracy appropriate for banks by reducing the risk of credit default. Banks could identify more candidates that will truly experience financial distress in the next two years. Correctly identifying individuals with high risks of default may prevent banks from financial losses since the bank can get the loan back in time with fewer people defaulting. Correctly identifying individuals with low risks of default may increase the bank's profit since the bank can be more confident in granting them more loans being assured that they have the financial capability of repaying. The bank can generate more revenue from the interest it charges more applicants with.

The optimization of FPR and FNR benefits young credit applicants such that they will not be treated disparately and denied a loan based on their age Ceteris paribus.

C. Would you be comfortable deploying this ADS in the public sector, or the industry? Why so or why not?

We would answer this question based on our analysis of the ADS. Since our ADS is fair as it does not contain any sensitive features and has a reasonable score on the fairness metric, we will say that it will be comfortable to deploy this ADS to the public. A fair system will reach its original goal of helping banks make the best decisions and benefit society. Neither the privileged nor the unprivileged will suffer from the disadvantage of the unfair result. The bank will gain its reputation and minimize the financial loss from selling lending bonds to people that won't be able to pay back.

However, before deploying the ADS into the public sector, we need to ensure that the ADS is robust enough to resist linkage attacks and be protective of people's privacy. Although the data is already anonymized as there are no personal identifiers such as names, SSNs, and phone numbers, the attackers may still attack people's privacy and identify individuals from the dataset using auxiliary information. Various industries may also utilize the data for a different purpose that may harm the individuals in the data. Therefore, there should not be more auxiliary data, and the industry that utilizes the data and the ADS should be trustable.

D. What improvements do you recommend to the data collection, processing, or analysis methodology?

Based on the current understanding of the data, I believe that the data collection procedure does not protect borrowers' information well enough, and it is not robust enough to protect people from privacy attacks. I would recommend using randomized response and differential privacy to transform the original dataset into data containing randomly generated data that will prevent others from recognizing each individual. The ADS can use DP synthetic data in independent attribute mode or correlated attribute mode that creates data similar to the distribution of the original data.

Moreover, we believe that ADS should follow ethnic principles and have respect for individuals. The ADS should receive informed consent from data contributors before they publish the data in competition for model training.

We would also recommend utilizing more metrics other than merely accuracy score to measure the performance of the model built within the system as the outcome seems to be highly unbalanced. The base rate of delinquency after preprocessing is 0.07169803012065878, which means a classifier that assigns 0 to each entry will achieve a 93% accuracy, we will suggest the ADS to compare false positive rates and false negative rates as we did in the above questions to evaluate the model.

Reference:

The ADS we analyzed in this report and corresponding datasets can be found at: https://www.kaggle.com/code/simonpfish/comp-stats-group-data-project-final