Erik Truong, Justin Tong
Final Project Report
Professor George Wood

Give Me Some Credit - EDA, Logistic Regression, WOE

Background

The purpose of this ADS is to predict whether someone will pay back a loan within the next 2 years or not by measuring an individual's credit score. This is the primary and only goal of the ADS. Using a logistic regression model, the weight of evidence module, and the random forest classifier, the models assess the different features within the dataset and find the most crucial predictors while optimizing the predictive power of the model. The first column within the model is the target variable and the remaining 10 are the predictors. The target variable is whether an individual will repay a loan within 2 year or not, and the 10 predictors are measures of debt ratio, credit utilization, monthly income, and number of times an individual had an outstanding loan within given time frames.

Input and Output

The data consists of 11 different columns which contain SeriousDlqin2yrs which is a boolean of if the person experienced 90 days past due delinquency or worse, RevolvingUtilizationOfUnsecuredLines is a percentage which is the total balance on credit cards and personal lines of credit except real estate and no installment debt like car loans divided by the sum of credit limits, age is the age of the borrower which is an integer, NumberOfTime30-59DaysPastDueNotWorse is an integer where the number of times borrower has been 30-59 days past due but no worse in the last 2 years, DebtRatio is the percentage of monthly debt payments, alimony, and living costs divided by monthly gross income, MonthlyIncome is the monthly income of the borrower, NumberOfOpenCreditLinesAndLoans is the number of Open loans (installment like car loan or mortgage) and Lines of credit (e.g. credit cards), NumberOfTimes90DaysLate is the number of times borrower has been 90 days or more past due, NumberRealEstateLoansOrLines is the number of mortgage and real estate loans including home equity lines of credit, NumberOfTime60-89DaysPastDueNotWorse is the number of times borrower has been 60-89 days past due but no worse in the last 2 years, and NumberOfDependents is the number of dependents in family excluding themselves (spouse, children etc.). The features that carried the most importance in the final model of this ADS were NumberOfTimes90DaysLate, RevolvingUtilizationOfUnsecuredLines, NumberOfTime60-89DaysPastDueNotWorse, and NumberOfTime30-59DaysPastDueNotWorse.

In the dataset, MonthlyIncome has around 20% of its data missing and NumberOfDependent is missing around 2.5% of its data. To deal with the missing data for MonthlyIncome, a zero was added in its place because those with missing incomes had a trend of having a high debt ratio where borrowers could be inserted with zero under the assumption that

they had made trivial income for the month. The NumberOfDependent shows that for rows missing MonthlyIncome, NumberOfDependent is also blank. Using that pattern as well as logically thinking, those without a monthly income would likely not have dependents so all the empty data cells are replaced with 0.

The output data and goal of the data set is to find the best AUC score to predict whether or not someone will experience financial distress in the next two years. The AUC score is used to determine the performance of the model at distinguishing between positive and negative values. The ADS used in the model had an AUC score of 0.868 which is a very good score. The ADS is predicting whether or not someone will experience financial distress in the next two years with an 86.8% accuracy.

Implementation and Validation

Within the dataset, about 20% of the values under the "Monthly Income" column and about 2.6% of the values under "Number of Dependents" column are left blank, or were null. The creator of the ADS speculates that the monthly income values are left blank due to being trivial workers and their relatively high debt ratio. The individuals that left the number of dependents blank often had no dependents and were incidentally the same individuals who didn't fill in their monthly income. These people often had 0 dependents. As a result, both of the columns were imputed with values of 0. Imputing the values with 0 were also consistent with the range of the variables.

The ADS's implementation begins with exploratory data analysis to reveal that the dataset is imbalanced, which is solved through data processing techniques such as SMOTE and Tomek Links. The training data of certain variables are also very heavily right skewed, so it can be noted that outliers are a prominent part of the dataset. Following other analyses of the different predictors, a baseline logistic regression model is performed in order to assess the model's performance before any optimizations are performed. Weight of evidence is then used, which is a module frequently used to measure likelihood of default, assessing the amount of information that is given by each variable. Through the preprocessing methods of fine and course classification and implementing dummy variables, the variables are optimized and the model's performance is further improved.

The ADS's performance is assessed through the use of metrics such as: precision, recall, F1-score, and the ROC-AUC value. The p-values of each predictor is taken and a confusion matrix is displayed following each implementation of the logistic regression model. By dividing the initial dataset into training, validation, and testing sets, each model's result was tested against both the validation and the test set. By recording the probability of an individual being able to repay a loan within a 2 year period, whoever is in charge of deciding the probability threshold is able to "pass" or allow a loan to be given to an individual or denied a loan. This probability decision is the way the model is able to perform its stated goal.

Erik Truong, Justin Tong
Final Project Report
Professor George Wood

Outcomes

One of the ways we sought to improve this ADS is to improve the fairness of our model. To do this, we can create a metric frame that measures the sample size, selection rate, false negative rate, false positive rate, accuracy, average precision, and roc auc score for different age groups. The data that is inputted into the metric frame are the model predictions, the test data frame, and the sensitive feature is the age of the subject.

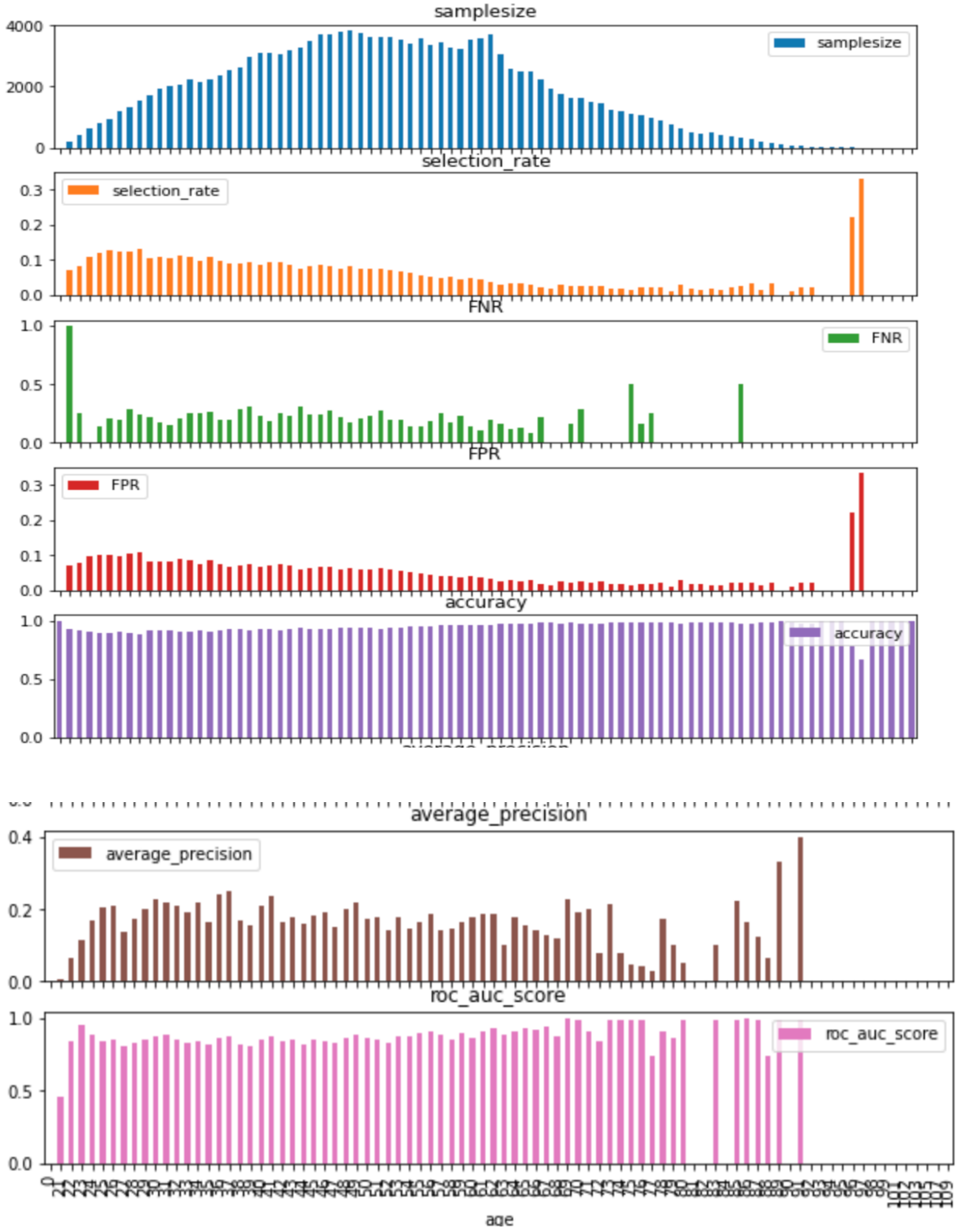
```
[[139364   7815]
 [   610   2211]]
```

This is the confusion matrix of the predictions and we can see of 150000 predictions, 141575 were predicted correctly with 139364 being true positive and 2211 being true negative. The rest of the confusion matrix tells us that 7815 are false positives and 610 are false-negative results.

```
samplesize           150000
selection_rate      0.06684
FNR                0.216235
FPR                0.053099
accuracy           0.943833
average_precision  0.176908
roc_auc_score      0.865333
dtype: object
```

Looking at this summary of the model, we can see the rates of the results from the previous confusion matrix. The false negative rate is higher than the false positive rate, meaning that this test is more specific and less sensitive. We also can see that although accuracy is high, the precision is not, meaning that the model grouped predictions together but does not necessarily mean that those predictions were correct. The ROC AUC Score is relatively high showing its proficiency at separating different classes within the model.

Erik Truong, Justin Tong
Final Project Report
Professor George Wood



Here are the results of the model graphed out with the x axis at the bottom being age. From sample size and selection rate, we can see that the majority of the people are in their 40s

with very few people in their 90s. We can also see for the false negative and positive rates, they are opposites of one another. The false negative rate has a peak during the early 20s and then is level after to be around 0.25. The false positive rate decreases starting at 0.1 then a major spike around 98 and 99 years sold. This can tell us that the two age extremes are predicted incorrectly by the model more than any other age group. The accuracy is almost uniform for all age groups with a slight decrease around 97 years old. The average precision hovers around 0.2 and 0.1 with it being lower at early ages like 21 and 22 while increasing in the late 80s and early 90s. The ROC AUC score stays high throughout age groups except for 21 year olds at around 0.4 while every other age group is around 0.8 or 0.9.

```
samplesize              3836
selection_rate      0.333333
FNR                      1.0
FPR                 0.333333
accuracy            0.333333
average_precision   0.394536
roc_auc_score       0.529557
dtype: object
```

This is the summary for the difference between metrics between subgroups for the group with the highest and lowest metrics. From the false negative rate, we can conclude that for the max group the false negative rate was 0 and for the other it was 1. This is probably due to one group having a small sample size as there are few people in the data set that represent very old ages.

Erik Truong, Justin Tong
Final Project Report
Professor George Wood

```
samplesize                149999
selection_rate          0.266493
FNR                     0.783765
FPR                     0.280235
accuracy                0.277167
average_precision       0.223092
roc_auc_score           0.401047
dtype: object
```

This shows the difference for the overall values of the metric. We can see how a few values such as the false negative rate and false positive rates increased while ROC AUC score and accuracy has decreased. This shows the variability between all the different age groups in the dataset.

The dataset only contains one sensitive attribute, which is age. Historically, younger people have a harder time to get approved for a loan. In order to calculate the fairness of this model using statistical parity and disparate impact as the fairness measures, age was divided into 2 categories. Individuals who were over the age of 45 were considered part of the privileged group and those under 45 were considered the unprivileged group. 45 was the number decided as the threshold because this is considered the start of being "middle aged." While the median of the dataset is 52, that is not necessarily an age someone would be "young." That being said, the threshold is subjective and can vary depending on the person.

```
Statistical Parity:  4.7097065459703045e-06
Disparate Impact:  1.001058527563299
```

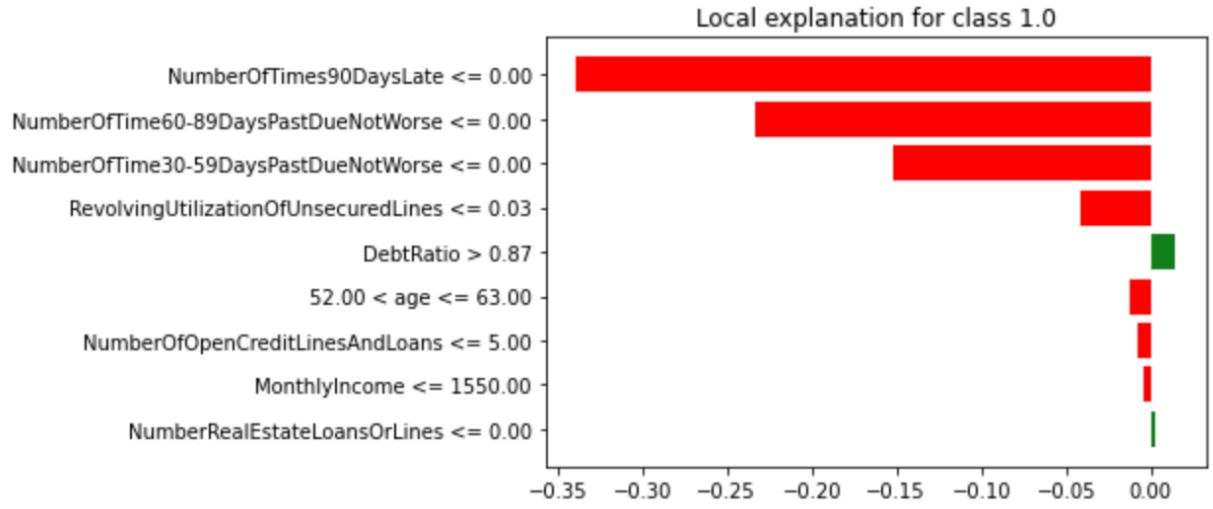
```
Statistical Parity:  [ 7.60615532e-06  6.26484269e-05  1.89448780e-03 -2.70147112e-04
 -1.67104674e-03]
```

```
Disparate Impact:  [1.00034473 1.00284164 1.08825299 0.98780521 0.92604806]
```

We analyzed the pre-existing bias of the dataset because the various optimizations that were performed on the final ADS were not related to age, indicating the 60-90+ days of an outstanding loan was the most impactful predictor. However, pre-existing biases still could not be discounted from the original dataset. After indicating which groups were the privileged and unprivileged groups, the dataset that was used to train the dataset was first broken down into 5

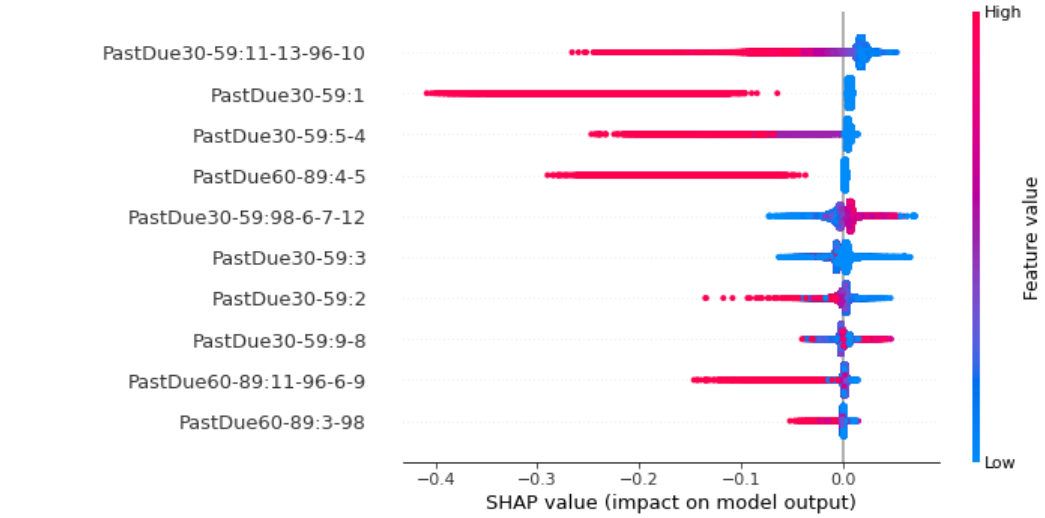
separate groups, each of size 30,000 to match the size of the predictions. Comparing the training data against the prediction data, we can see that there is negligible statistical parity that is present between the predictions and the dataset. However, we can see that the disparate impact rises over the different subsets of 30,000 in the dataset. Within the first 30,000, we can see that the data actually favored the unprivileged groups, but it is very small. Then the value progressively drops in the following 4 iterations. The average disparate impact of the dataset is 1.001. This value shows that there is a fairly noticeable amount of disparate impact in the dataset, indicating older individuals were more likely to receive positive/favorable outcomes over the younger, underprivileged groups. That being said, I believe this is an acceptable amount of disparate impact, as there is the high likelihood of distorting the data if we were to process either the predictions or the original dataset. Considering the goal of the ADS, it is to try to accurately predict the likelihood of whether or not an individual is able to repay a loan within 2 years. The resulting imbalance in the ADS could potentially be explained by the correlation between the different features, so the bias that is present could likely just be a consequence of other factors impacting age-fairness.

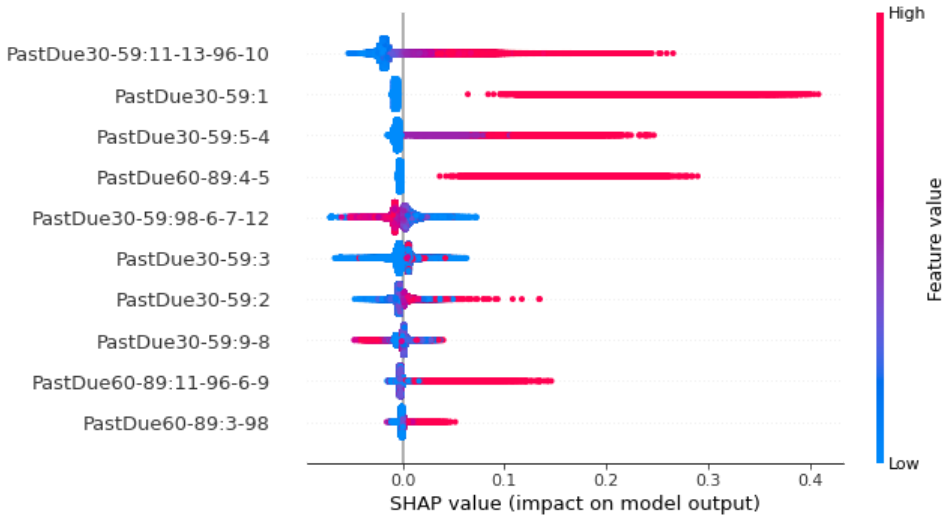
We also created a LIME representation for the data set that showed the explanations for each of the variables. As a note, this particular dataset did not have any categorical variables.



From this graph we can see the dependent variables in this data set. The variables that are best predictors are if someone has a history of not paying back their debts. If they have this history, the model is more likely to predict them to have financial distress in the next two years. There are only a few factors that deter the model from predicting someone to have financial distress, such as debt ratio and number of real estate loans or lines. Number or real estate loans is ambiguous because it could be from a real estate investor rather than someone that is finding an

initial place to live. This is also highly dependent on the current housing market, as if the housing market is bad, then an investor may face financial difficulties, which can be an explanation for why this is the most insignificant variable. For the debt ratio, if someone has a debt ratio lower than 0.87, they are less likely to face financial stress. The weights of the variables are explained later in the report with SHAP values.





Attached above are the shap values for the most impactful predictors of the dataset. We know these are the best predictors of the dataset thanks to the exploratory data analysis performed by the original coder of the ADS. By plotting the distributions before and after different baselines models and through the use of Weight of Evidence, the best predictors were identified and optimized prior to being refitted into the random forest model. By plotting these shap values, the decisions made by the decision tree are made easy to interpret by showing how the presence of this predictor could affect the resulting prediction.

The two different plots attached above are essentially the same, but one contains the negative shap values and the other are the positive shap values. Considering the magnitude of

each predictor is the same, just with different signs, they can essentially make or break an individual's ability to be approved for a loan. In response to the robustness of the model, with the help of the different black box models like SHAP and LIME, we can see that the model is built to be resilient to extreme values/outliers.

Summary

Yes, we believe that the data was appropriate for the ADS. The predictors that were provided involved very relevant data points that are vital for determining a person's ability to pay back a loan. Information pertaining to the number of loans a person has, the number of times a person had an outstanding loan within specified time limits, monthly income relative to the number of dependents, revolving utilization of unsecured lines, and debt ratio. Each of these predictors are important in a bank's decision-making process. The number of loans a person has reveals the number of current loans a person has which includes real estate loans. Regarding debt ratio, monthly income, and number of dependents, these are all factors that affect a person's ability to repay a loan. In addition to the number of loans, each of these financial stressors and measurements are crucial in determining a person's financial burden. The more financial burden, the less likely they would be able to pay back the loan within the given time frame. With revolving utilization of unsecured lines, this metric is able to give further insight in how frequently the person uses an allotted credit limit. Under the assumption that they repay in a timely manner, the higher the ratio the better. Finally, the number of times a person had an outstanding loan between 1, 2, and 3 month time frames are great indicators of how frequently a person is late in repaying a loan. These predictors are necessary for the implementation of a dataset, with a single sensitive feature involving age, which can also give insight to the amount of experience an individual has with credit.

Regarding the implementation of the ADS itself, we would argue that the implementation is robust, accurate (in its most literal sense), and fair. To begin, the author of this ADS implemented 2 different classification models, each with their own optimizations. A logistic regression using weight of evidence coursing and a random forest classifier, both utilizing pipelines to further optimize each model. The random forest classifier was the one that was ultimately submitted, and those are particularly robust to outliers considering the number of subset sampling and bootstrapping performed on the data, so this ADS is robust. In terms of accuracy, we utilized Fairlearn and BinaryLabelDataset to measure statistical parity, disparate impact, accuracy, precision, and false positive/negative rates. Regarding the accuracy of the dataset, we argue that the accuracy measure of the model itself was amazing. However, the precision of the model was terrible. With an average precision of .223, if the bank was trying to be very safe with who they give loans to, then this would be a wonderful option. Finally, regarding the fairness of the model, there was almost no statistical parity in the dataset after comparing the predictions, and the disparate impact was above .93. These metrics indicate that each individual within their own age group was fairly judged and the likelihood of being

approved was the same throughout. The measure of disparate impact, with age being the sensitive feature, was within an acceptable range. By acceptable range, I mean there could be other factors that could contribute to this disproportion between age groups that is not entirely decided based on age. A younger individual would not have as much time to collect work experience, hold a better paying position within a company, or have as much experience with credit, so these other predictors contribute to the slight bias in age groups. Thus, this is a robust, accurate (strictly speaking about the accuracy metric), and fair ADS.

In the public sector, I would be comfortable deploying this ADS because of the low precision of the ADS. Although the accuracy is high, the precision is not great, meaning that groups are identified correctly but their placements may be incorrect. A low precision means that there are more false positives which in this case means someone that is labeled as someone that would default on their loans but in actuality they don't. Banks in the public sector are meant to serve the people and help small businesses and people with loans so they may be willing to take more risk in order to stimulate the economy. This ADS is reasonable in its predictions where applicants still have to pass many parameters as a screener, but this ADS would allow the public sector to move much quicker when issuing out loans and the model has good enough accuracy that there are hopes that the losses covered by the few that do default on their loans are covered by those that pay it back. In the industry, I would not be comfortable deploying this ADS because banks in the private sector are trying to maximize profits and minimize risk. Allowing an ADS with subpar precision would increase the risk to the banks for these loans. A private bank has a few options, with a few being to continue to refine the ADS to become more accurate or manually review every applicant that wants a loan and predict their likelihood to default on a loan with information they provide. Especially with stakeholders of large companies, many of them are willing to forgo potential profit for less risk, so having an ADS that increases risk at the potential for increased profit is not a viable option.

For one, the data providers could have given more information about how the data set was obtained. The only information that was given was historical data from 250,000 borrowers. The data could be collected from more reliable sources but we do not know. Another is that the data could have more predictors such as currently employed or education level that could potentially help make the model more accurate. Although the data is going to be much less abundant, there could be more samples for the older people asking for loans because they could potentially skew results because of their extreme values and meager amount of samples.

This ADS did very well with its processing and analysis method. For example, there were missing values for monthly income and number of dependents for some rows. To deal with this, it was reasoned that since those with missing monthly income had a high debt ratio, and those with a high debt ratio usually did not have a monthly income, then the missing values were replaced by 0s. For the missing number of dependents, they were usually missing with missing monthly income and logically, those without dependents would also have little to no monthly income.