Audit of an Automated Credit Card Approval System

Ethan [redacted], [redacted]

May 10th, 2021

Background

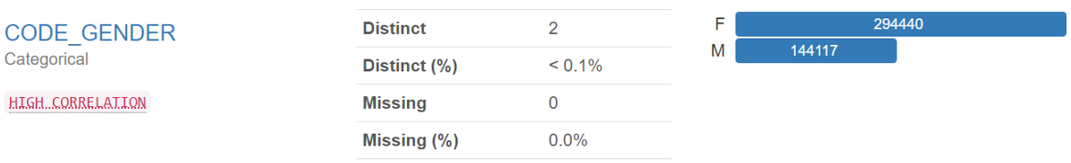
Automated decision systems (ADS) are largely used across banks and other private financial institutions for risk assessment. These systems combine information provided by individuals, alongside other personal data to quantify the likeliness of defaulting on a loan or a credit card. The inner workings of these systems are often intentionally obscured, and as a result, the general public has little transparency into how these systems are using their data. In this analysis, we will be evaluating one such ADS with a "nutritional label" to understand the fitness of its use for classifying individuals for credit card approval. The ADS, named "Credit Card Approval Prediction Using ML", can be found on Kaggle [1] and its creator is an individual who goes by the screen name [redacted]. The accompanying data set is from a Chinese bank and contains 21 columns of personal attributes and just under 450,000 rows of individuals.

While the primary goal of this ADS is to predict whether an individual will be approved for a credit card, it is also used to test several algorithms in an attempt to determine the most accurate metric for prediction. The approval status is based on the aforementioned personal attributes which includes age, gender, occupation, earning history, family status & size, and overall wealth. Because there is no universal standard to determine whether or not an individual is high risk, [redacted] liberally defines the threshold of high risk as individuals who have been overdue for more than 60 days. While exploring this threshold is not the primary objective of this analysis, its existence carries some ramifications for the ADS; Because the threshold was chosen without proper explanation or documentation, a conflict of interest arises where such a threshold might have been chosen to maximize model accuracy instead of fairness. A large number of individuals were also dropped from the analysis which could disproportionately bias the ADS towards accepting or rejecting certain groups of individuals over others. Over the course of this evaluation, we will look extensively at the author's intent in decision making and their choice of algorithm implementation to accurately assign a nutritional label.

Inputs and Outputs

While the data used in this analysis was initially collected by a Chinese bank, currently it's main repository is now on Kaggle. [redacted], the creator of the ADS, is also the initial uploader of the data and maintains a discussion about it on the platform. [redacted] claims to have been provided the data from a popular Chinese online course[2] with the same objective as the Kaggle competition. Unfortunately, not a lot of information is provided about how the data was collected, who collected it, or who originally requested the creation of the data set in the first place. As a result, the context surrounding the data remains a mystery.

One of the main important attributes in this data set is a binary classification of gender. Although there are no missing values, we see almost twice the amount of females as males.



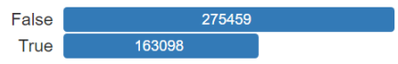
Assets such as property and car ownership are also included. These have no missing values and are binary, like gender. Only about a third of individuals own a car while two-thirds of individuals own property.

[1] https://www.kaggle.com/rikdifos/credit-card-approval-prediction-using-ml
[2] https://mp.weixin.qq.com/s/upjzuPg5AMIDsGxlpqnoCg

FLAG_OWN_CAR
Boolean

HIGH_CORRELATION

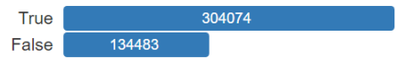
Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	428.4 KiB



FLAG_OWN_REALTY
Boolean

HIGH_CORRELATION

Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%

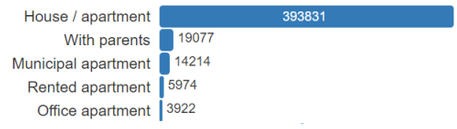


The data also includes housing types where over 75 percent of individuals live in a house/apartment.

NAME_HOUSING_TYPE
Categorical

HIGH_CORRELATION

Distinct	6
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	30.9 MiB

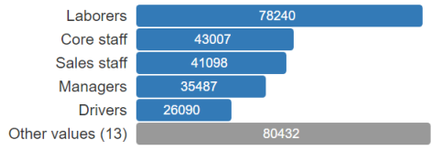


Occupation type is also included and is the largest source of missing data. The data is skewed to the left where most individuals are laborers. Core staff and sales staff come second and third respectively with roughly the same amount of individuals in both groups. Most individuals fall below 0, indicating that they have been employed for a given period of time.

OCCUPATION_TYPE
Categorical

HIGH_CORRELATION
MISSING

Distinct	18
Distinct (%)	< 0.1%
Missing	134203
Missing (%)	30.6%
Memory size	23.7 MiB

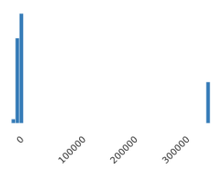


Days employed is a numeric value denoting how many days an individual has been at their job. The value counts backwards from the day started so negative values with a larger magnitude indicate a longer period of time worked while any positive value indicates how many days an individual are unemployed. We can see values here that are extremely high and corresponding to almost 800-1000 years. This is because pensioners have a filler constant in this column which is why the dates are so high.

DAYS_EMPLOYED
Real number (ℝ)

Distinct	9406
Distinct (%)	2.1%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	60563.67533

Minimum	-17531
Maximum	365243
Zeros	0
Zeros (%)	0.0%
Negative	363228
Negative (%)	82.8%
Memory size	3.3 MiB



Income was measured as total annual income and income type is the income category. On the chart below the income total is divided by 1 million. Income category is a string denoting where the money is coming from. The average that individuals make is around 190,00 USD with a minimum of just above 26,000 and a maximum of 6.75 million.

AMT_INCOME_TOTAL
Real number (ℝ≥0)

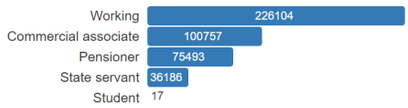
Distinct	866
Distinct (%)	0.2%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	187524.286

Minimum	26100
Maximum	6750000
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	3.3 MiB

NAME_INCOME_TYPE
Categorical

HIGH CORRELATION

Distinct	5
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	28.4 MiB

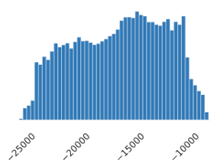


Date of birth for individuals is provided as a negative integer indicating the days since their data of birth. The larger the magnitude of this integer, the older the individual is. The data is skewed slightly towards younger individuals with the youngest being 20.5 and the oldest being just over 69.

DAYS_BIRTH
Real number (ℝ)

Distinct	16379
Distinct (%)	3.7%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	-15997.90465

Minimum	-25201
Maximum	-7489
Zeros	0
Zeros (%)	0.0%
Negative	438557
Negative (%)	100.0%
Memory size	3.3 MiB



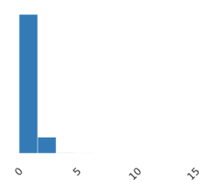
Data on families was also provided. Children Count measures how many children an individual has, and CNT FAM Members measures the total family size. Most individuals do not have children with almost 70 percent of respondents indicating they were childless. Family status denotes the marriage status of an individual and more than 60 percent of individuals indicated that they were married.

CNT_CHILDREN
Real number (ℝ≥0)

ZEROS

Distinct	12
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	0.4273902822

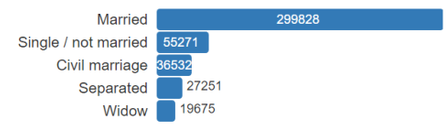
Minimum	0
Maximum	19
Zeros	304071
Zeros (%)	69.3%
Negative	0
Negative (%)	0.0%
Memory size	3.3 MiB



NAME_FAMILY_STATUS
Categorical

HIGH CORRELATION

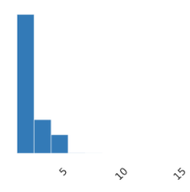
Distinct	5
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	27.7 MiB



CNT_FAM_MEMBERS
Real number (ℝ≥0)

Distinct	13
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	2.19446503

Minimum	1
Maximum	20
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	3.3 MiB

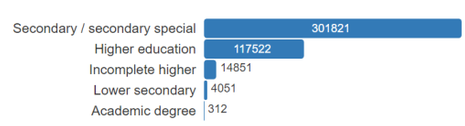


Education level of each individual is provided with most individuals having at most a secondary education. About 20 percent of individuals also have a completed higher education.

NAME_EDUCATION_TYPE
Categorical

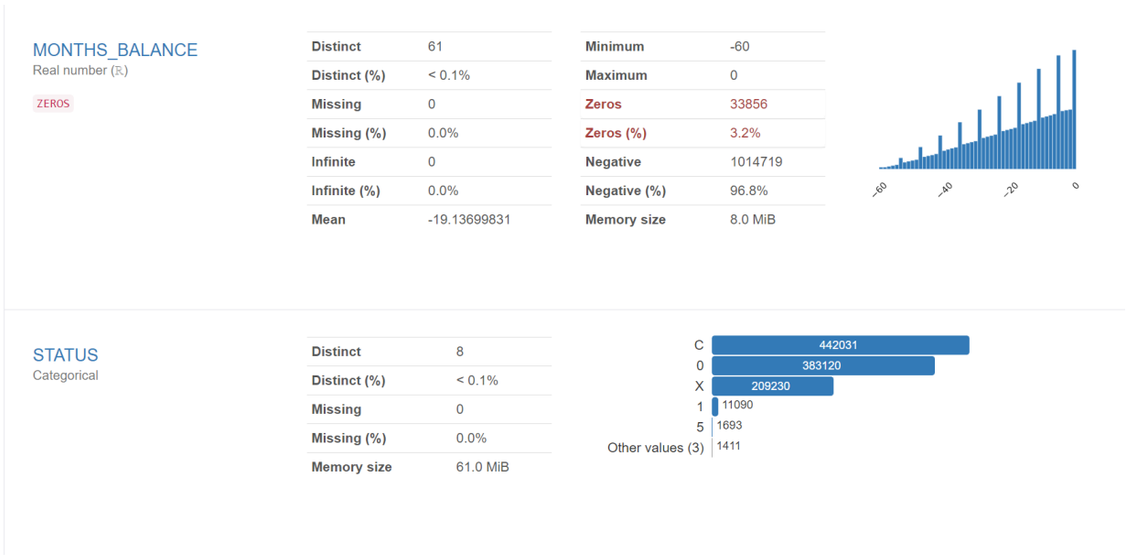
HIGH CORRELATION

Distinct	5
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	34.3 MiB

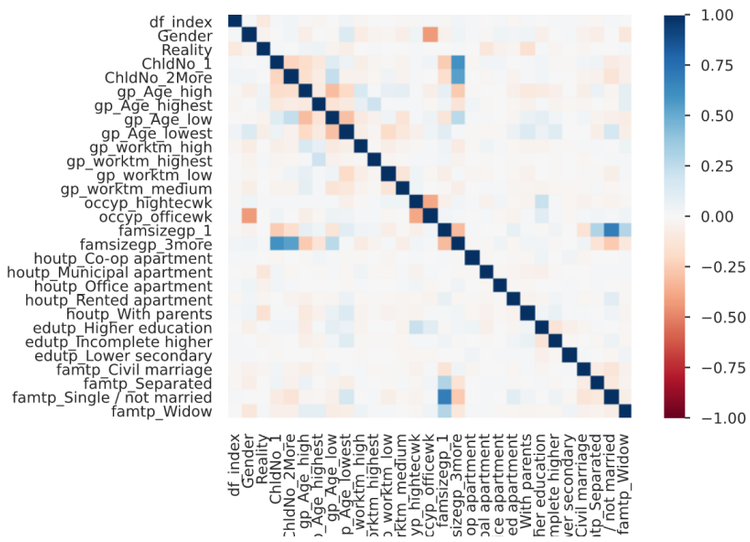


Month Balance is the month that the data was recorded where the month of the extracted data is the

starting point. For example, 0 would be the current month, -1 is the previous month. This accompanies Status which is the outcome of interest for this ADS where the classification is; 0: 1-29 days past due 1: 30-59 days past due, 2: 60-89 days overdue, 3: 90-119 days overdue, 4: 120-149 days overdue, 5: Overdue or bad debts, write-offs for more than 150 days, C: paid off that month, and X: No loan for the month.



The pairwise correlation between variables is included below. This plot was generated after the data was processed and some variables were one-hot-encoded to capture the entire scope of the data. We see correlations that we would expect, such as family size and marriage status, but also a seemingly strong correlation between having an office occupation and gender. The output of this ADS is a binary "yes" or "no" indicating whether or not an individual has been approved for a new credit card. Accuracy measurements for each of the classification algorithms are also included in the main output.



Implementation and Validation

The initial data is provided in two parts: Application Record and Credit Record. The creator of the ADS uses Months Balance from Credit Record to create a variable to indicate the date of the "beginning of account." The Status column is then used to designate a threshold for individuals who are high risk of defaulting and not getting approved. In this case, 2 months was chosen as the arbitrary cutoff for being likely to default on the credit card. This means that any individual 2 months overdue or more were treated as individuals who would not be approved for a credit card by a bank. The total percentage of individuals who fell in this category came to 1.5 percent. The two CSVs are then merged by ID to create one data set of about 36,000 individuals. Next, columns are renamed and all null values are dropped which brought the data down to a total of 25,000 valid observations. After the dropping of the null values, the binary categorical variables (like gender, house/no house, car/no car) were then changed to 0 and 1 while continuous variables (like number of children, age and annual income) are binned. Synthetic Minority Over-Sampling Technique (SMOTE) was then used to overcome sample imbalance within the data and finally, all non-transformed variables with low IVs were dropped and the data was split into training and test sets.

SMOTE is the key technique implemented by the author of the ADS to reduce imbalance and attempt a fair implementation of the various algorithms. SMOTE over samples observations from the minority class of the target variable and creates as many synthetic observations as necessary to obtain balance on the target variable. Before the null value drop, the data has 67 percent female and 33 percent male, however, after the drop that number changes to 62 percent female and 38 percent male. After the SMOTE balancing, the female majority drops to 59 percent which is closer to the ideal 50/50 gender split. SMOTE is entirely successful in creating perfect balance on the target variable with 24,712 observations for each outcome for a total of just under 50,000 total observations.

After pre-processing, Information Value (IV) was calculated to measure the predictive power of various independent variables in relation to the binary dependent variable. IV is calculated by variable, without taking into account other predictors. The higher the IV value is, the higher the predictive power is and if the IV value is too low, that variable would most likely not accurately describe the target variable so it should be ignored. Below is a collection of the IV values calculated for all the major variables:

	variable	IV
10	agegp	0.0659351
8	famtp	0.0431371
11	worktmgp	0.0402215
3	Reality	0.0274407
1	Gender	0.0252035
7	edutp	0.0103618
9	houtp	0.0073275
17	famsize	0.00615614
16	occyp	0.00482047
5	incgp	0.002422
13	wkphone	0.00204243
4	ChldNo	0.00112145
14	phone	0.00054805
6	inctp	5.1593e-05
15	email	1.73436e-05
2	Car	4.54248e-06

Looking at the IV values, the ADS author concludes that Gender is an important predictor of good credit and chance of credit card approval. Gender ranks even higher than education, family size, having a house, income group etc. Age and family type (single, married etc.) both have bigger impacts. Even though these IV values are not used explicitly, they are used to evaluate which variables are of importance. Based on this analysis, [redacted] drops the variable to flag whether someone has a car or an email.

After all the pre-processing is complete, the training and test sets are put into each of the models. Logistic Regression, Decision Trees, SVM, LightGBM, XGBoost and CatBoost (a package to boost gradients in decision trees) are the algorithms used for classification. The implementation and results for these are presented sequentially to showcase increasing complexity and accuracy. CatBoost is the main algorithm that we will be analyzing since it is the final model presented with the best accuracy score. For the purposes of this ADS analysis, we remove all the other models from the accompanying python notebook. CatBoost is specifically used to counter overfitting and potential gradient boosting biases which makes it more robust relative to other decision tree algorithms. Models validation was done on the test set and accuracy was the only metric calculated and returned. The goal of successfully classifying individuals was achieved in the eyes of the creator, however, there is no real discussion about whether or not the ADS is fair or robust across various populations. Additionally, the target values were made up by the author themselves. Since only providing accuracy is not a robust method for validation, we will implement *aif360* and SHAP to evaluate fairness alongside other custom metrics during the pre and post-processing periods.

Outcomes

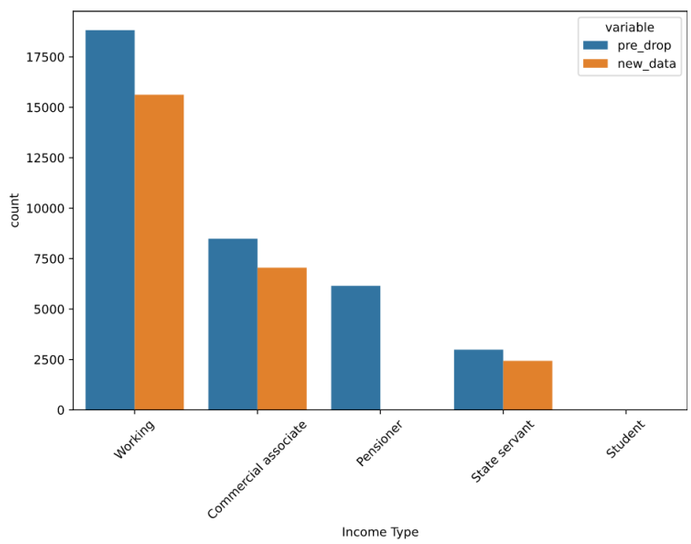
In this analysis, gender is our main variable of interest. Out of all the models presented by the notebook author, CatBoost was the most accurate and fair with an overall accuracy of 95 percent and a a disparate impact of 1.18 leaning towards the female population.

Before looking at outcomes, we decided to look into the pre-processing steps to ensure that all possible bias was taken into account before making fundamental changes to the data. We found that the dropping of null values was a big mistake that fundamentally changed the outcome of the ADS for certain sub-populations. As discussed in the input section, almost every "null" value present in the data set came from the "Occupation Type" column and when these observations were dropped, no attention was paid to the potential distribution changes in the sample. As part of our analysis, we performed the KL Test for

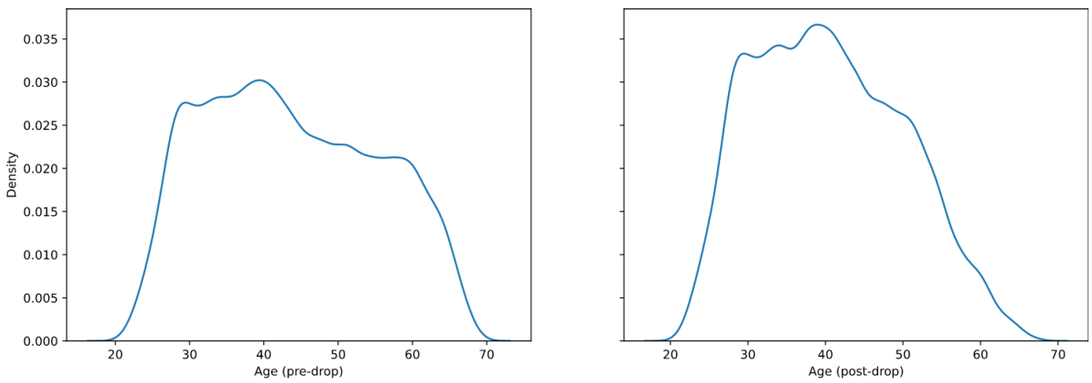
divergence to see possible changes in the distribution of individual characteristics. This table is presented below:

	variable	KL_divergence
0	occyp	0.000
1	FLAG_MOBIL	0.000
2	target	0.000
3	dep_value	0.000
4	phone	0.000
5	begin_month	0.000
6	Reality	0.001
7	email	0.001
8	edutp	0.001
9	houtp	0.001
10	Car	0.003
11	Gender	0.005
12	famtp	0.005
13	wkphone	0.006
14	famsize	0.007
15	ChldNo	0.008
16	inc	0.019
17	inctp	0.181
18	DAYS_EMPLOYED	0.272
19	DAYS_BIRTH	0.309
20	ID	0.372

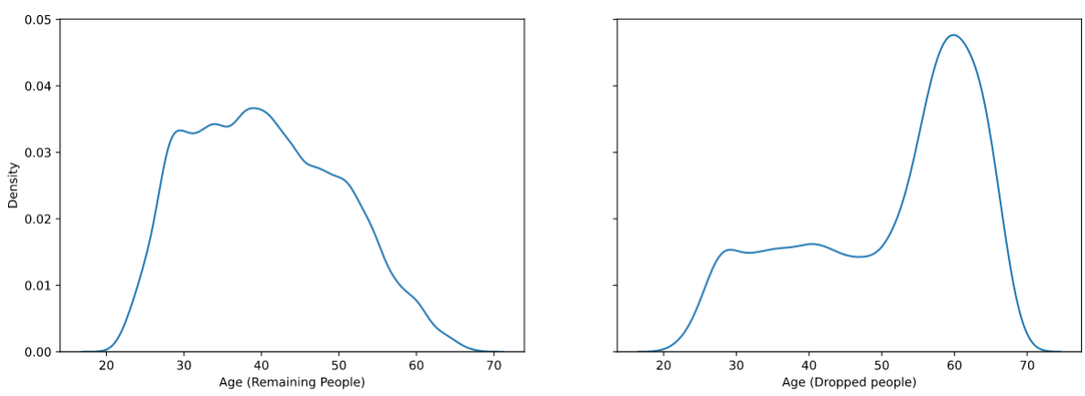
As we would anticipate, since only null values were dropped, there is no difference in occupation distribution. Most variables only had a slight change which is ideally what we would want. Even gender only has a value of 0.005 which is minimal. However, the employment time and age of individuals seems to have changed drastically with values of 0.272 and 0.309 respectively. Income type also drastically changed with a test score of 0.181. We decided to dive into the data to figure out what theses differences meant and how it impacted the distribution of individuals being fed to the various algorithms. First, we decided to look at the income type to see if there were any fundamental changes in the types of individuals who were present before and after the null values were dropped.



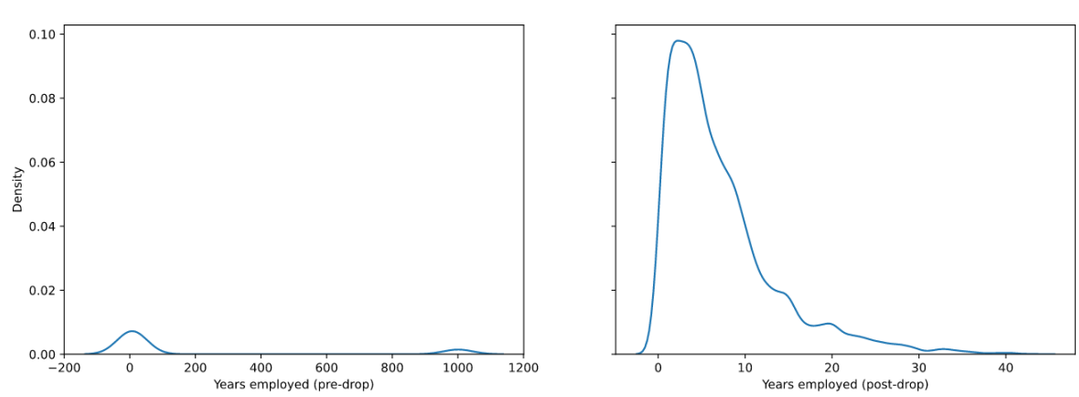
As we can see in the figure, Working, Commercial Associate, State Servant, and Student remained largely unchanged, however, Pensioners almost completely disappeared. To double check this disappearance of presumably older individuals, we plotted the distribution of ages across both data sets:



We see that overall, the distribution has shifted left in favor of younger individuals which means the ADS is fundamentally leaving out a subset of the population that was present before the pre-processing. We then decided to look at the distribution of individuals who were dropped:



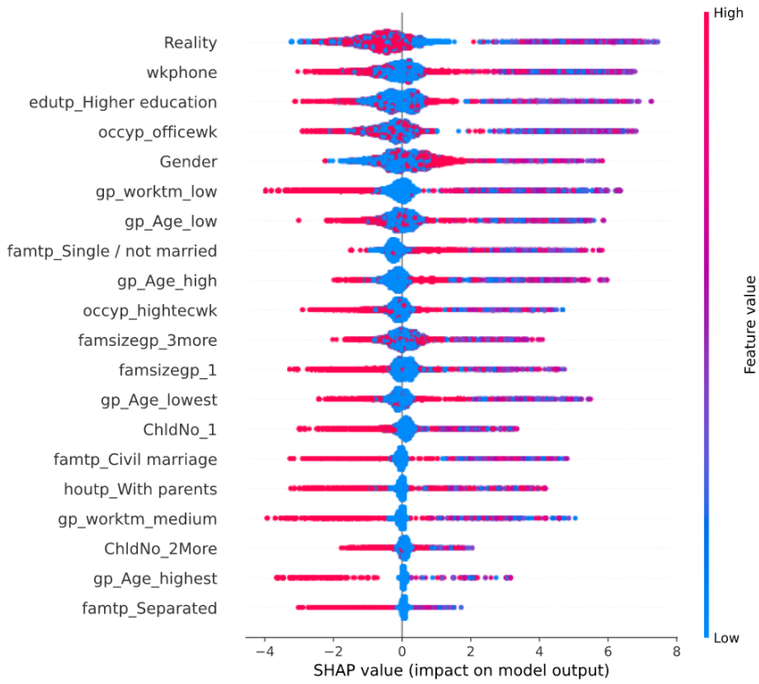
Here it is evident to see that a majority of the dropped individuals were part of the older population. It seems that the original data did not list occupation type for Pensioners. These are the individuals who had a filler variable for time worked which we can see were completely removed from the data:



As expected, all other distributions pre and post drop looked extremely similar which is what we would expect based off the KL test values. So, fundamentally the sample was changed to not include Pensioners and by extension, a large portion of the older population. Another thing to note is that by carelessly dropping all null values and thus the people with no listed occupation, the author may be disadvantaging those who may not be able to work for various reasons and thus have no recorded occupation. There is no mention of this in the ADS and the author just assumes that the sample distributions were the same which would most certainly cause bias during ADS deployment.

The main fairness measure we used here was disparate impact which is the ratio of receiving a positive outcome based on being in a particular group. In this case, we set the unprivileged group as the female population. We believe this is an appropriate metric due to the high stakes of the ADS and the existence of gender as a protected attribute in our data. It is important that the outcomes be similar for both genders since Credit approval is no grounds for differential treatment and is a domain that deserves demographic parity. We calculate the disparate impact to be $\sim$1.18 which indicates that females are more likely than males to receive approval for a credit card. We also calculated the mean difference in outcomes between females and males which was 0.085.

We followed this up by calculating the SHAP values for the CatBoost model. The Shapley value summary plot for the most important variables is shown here:



On the X-Axis we see the impact on the model. Positive values indicate more likelihood of being rejected and negative values mean better chances of approval. The redder values correspond to higher feature values – so in the case of binary variables, this translates to those people who had a 1 under a specific column. With this in mind, we see that owning "realty" is a good indicator to the CatBoost model that a person is worthy of "credit approval." We can see that for almost all variables, a positive value (like having higher education or being in a civil marriage) is associated with an increased likelihood of not being rejected by the ADS. This is what we would expect to see. We do see that there is a segregation of males and females based on their SHAP values such that males are less likely to get approved for Credit Cards. All in all, we admit that the SMOTE balancing done by the author to balance the data ends up lessening the bias in the data and brings the ADS to an impressive 95% accuracy. But the data was too unreliable (and mostly synthetic) which in addition to the sharp dropping of values probably made the ADS not fit for production.

Summary

In this analysis, we looked at an automated credit card approval system using data from a bank in China. While the data had many relevant individual characteristics that would help create such an ADS, the numerous issues with the data make it sub-optimal for creating a robust and viable system. The creator attempted to correct for some of the major problems such as the large gender imbalance as well as the presence of missing values, however, these attempts fell short of creating a credit approval system.

While only accuracy scores were initially provided, the additional metrics that we calculated indicated that the some sub-populations were unfairly disadvantaged or excluded from the analysis all together. This form of careless accuracy-chasing could be quite harmful if let loose in the real world. We would not feel comfortable deploying this in the public sector nor the industry as the data that this ADS is trained on is too imbalanced and not representative of the overall population.

As explained in the rest of the analysis, the greatest weakness in this ADS is the heavily imbalanced data which it was trained on and thousands of individuals were further lost from the data during pre-processing which only left individuals who were "employed" or "employable" for the training of the ADS. It would be no surprise if the ADS showed differential treatment against differently-abled or older populations. Apart from using a more balanced data set, it is recommended that more attention be paid to analyse who the ADS serves and who gets left out. This Kaggle notebook (pitched as "Credit Approval using ML") is currently the most highly voted implementation in the Credit Card Approval challenge and gets viewed very often. We know very well that ML today has a halo on its head for those in the industry and seems very alluring to those who don't understand it. It is our hope that going forward, authors like [redacted] make an effort to include some thorough impact-analysis along with their attempts at fair ADS.