Project Report

Bernice Wu, Ding Jin

April 2021

Background

Context

Credit score cards are widely used in commercial banking. As a part of risk management, personal information of the clients are gathered and analyzed to predict of probability of default and the allowance for borrowing. Commercial banks are capable of bridging the gap between personal information and potential risk of clients by developing algorithms such as machine learning.

Goal

Build a machine learning model to predict the credit level of a particular client so that commercial banks could distribute loans accordingly and minimize their risk. In order to reduce the overall risk level, the bias of the data (pre-existing, technical and emergent) should be mitigated. In this way, we could generate more accurate outcomes that precisely predict the credibility of customers. A client is supposed to be labeled as "good" or "bad" based on the algorithm developers' discretion. Also, there should be mitigation methods to tame the unbalanced data in this task. So, there may be some trade off between accuracy and fairness.

Input and Output

Description of Dataset

The dataset is generated from one of the Kaggle competitions named Credit Card Approval Prediction. The main focus of this competition is to train the best model for predicting the credibility of one specific client.

The dataset consists of two csv files, naming 'credit_record' and 'application_record' respectively. 'credit_record.csv' keeps track of users' behavior of the credit card and 'application_record.csv' is the dataset that records the appliers personal information which is considered as the set of features for predicting. 'credit_record' contains 1,048,575 observations and 3 columns that represent 3 features of one client, while 'application_record' contains 438,557 observations and 18 columns of features. These two datasets could be merged by customer id.

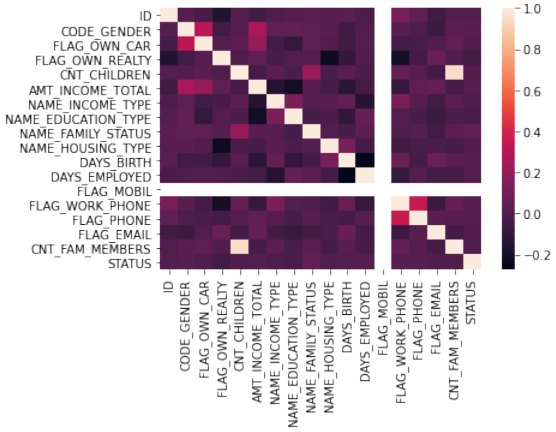
Feature

For 'credit_record' dataset, it contains the credit level information by tracking the clients' monthly movement with 2 features: month_balance and the status. For 'application_record', 17 features are used to evaluate the credit level of the clients, including gender, family status, educational background, financial status and career. Our choice of sensitive attribute is gender. ADS assume that the privileged group is male and unprivileged group is female.

Feature Table		
Feature Name	**Explanation**	**Data Type**
ID	Client number	int64
CODE_GENDER	Gender	Object
FLAG_OWN_CAR	Is there a car	Obejct
FLAG_OWN_REALTY	Is there a property	Object
CNT_CHILDREN	Number of children	int64
AMT_INCOME_TOTAL	Annual income	float64
NAME_INCOME_TYPE	Income category	Object
NAME_EDUCATION_TYPE	Education level	Object
NAME_FAMILY_STATUS	Marital status	Object
NAME_HOUSING_TYPE	Way of living	Object
DAYS_BIRTH	Birthday	int64
DAYS_EMPLOYED	Start date of employment	int64
FLAG_MOBIL	Is there a mobile phone	int64
FLAG_WORK_PHONE	Is there a work phone	int64
FLAG_PHONE	Is there a phone	int64
FLAG_EMAIL	Is there an email	int64
OCCUPATION_TYPE	Occupation	Object
CNT_FAM_MEMBERS	Family size	float64

Correlation between features

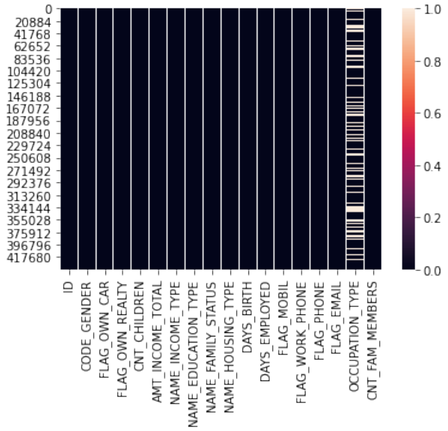
According to the heatmap below, ADS could get a general idea that most features are uncorrelated. While gender and car status have relatively high correlation, which suggests that there might be disparity in the property of car between male and female groups. There are also several pairs of features which shares apparent strong correlation: total income and income type, total income and education type, date of birth and days of employment, number of children and number of family members. Among them, the correlation between number of children and number of family members almost reaches 1, suggesting that ADS could then use one of them as a feature to predict. ADS could also see the blank stripe in the feature mobile phone status, which indicates that every customer in this dataset has a mobile phone, so it is uncorrelated with all the other features.



Correlation heatmap

Missing values

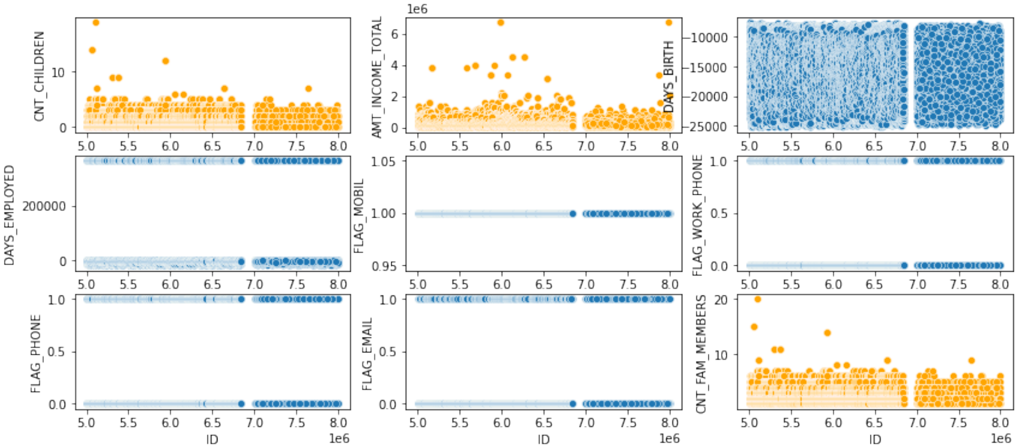
From the heatmap, ADS could see that there are multiple white strips in the column 'OCCUPATION_TYPE', which informs us that there are numerous missing values here and ADS need to either drop or impute the missing data.



missing values heatmap

Distribution of the input

Based on the plots, ADS could directly see the distribution of these attributes. Among the 9 features, ADS could observe that 'CNT_CHILDREN', 'AMT_INCOME_TOTAL', 'CNT_FAM_MEMBERS', 'DAYS_BIRTH' are continuous variables while the other features are categorical variables. In addition, there are some outliers in 'CNT_CHILDREN', 'AMT_INCOME_TOTAL' and 'CNT_FAM_MEMBERS' features, which needs to be further pre-processed before fitting in the algorithm.



missing values heatmap

Output

The output of this predictive algorithms is status of credibility, which is a binary variable with two potential values, 0 and 1. Status 0 represents that the customer is in good credit while status 1 suggests the bad credibility of the customer.

Implementation and Validation

Data cleaning and pre-processing

The ADS applied the data pre-processing before predicting based on the features information and graphical illustrations above. Basically, the author firstly dropped the features that do not have predicting value or have too many missing values to have prediction power. Then he evaluated the importance of the remaining features and found that all of them are significant for prediction. So, they could not be dropped at this moment, instead, they need to be converted into numeric variables in order to become appropriate features for prediction. In addition, the author performed a value count on credit status feature to check if there are oversampled data. The data here is highly imbalanced since 99.7% of customers are in 0 class, while only 0.3% of customers are in 1 class. Therefore, the author adjusted them by employing the oversample method from sklearn package before feeding them into a classifier, since otherwise the classifier will train itself to be more sensitive to detecting the majority class and less sensitive to the minority class.

High level information about the implementation of the system

After finishing data pre-processing, the author then fit all the selected features to six different classifiers, which are LogisticRegression, KNeighborsClassifier, SVC, DecisionTreeClassifier, RandomForestClassifier, XGBClassifier, and compare them in order to find the best classifier.

Logistic Regression is an elementary way to predict the binary class, which uses a logistic function to model a binary dependent variable.

KNeighbors classifier classifies an object based on the plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors. However, since this algorithm relies on distance for classification, if the features represent different physical units or come in similar scales then normalizing the training data can decrease its accuracy dramatically.

SVC classifier is one of the most robust prediction methods, being based on statistical learning frameworks or VC theory. Given a set of training examples, each marked as belonging to one of two categories, an SVC training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier.

Decision Tree Classifier uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves), which is also a robust prediction method.

Random Forest Tree classifier employs a multitude of decision trees at training time and therefore could tame the decision trees' habit of overfitting to their training set.

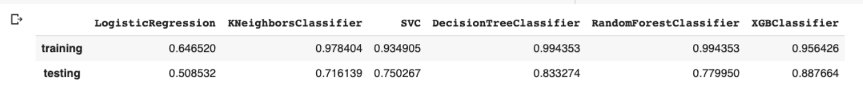
Lastly, the author employs the XGBClassifier, which is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. In prediction problems involving unstructured data (images, text, etc.) artificial neural networks tend to outperform all other algorithms or frameworks. However, when it comes to small-to-medium structured/tabular data, decision tree based algorithms are considered best-in-class right now.

Validation of ADS system

The author calculated the overall accuracy score for six classifiers and reported classifier precision for the best classifier. We found out that XGBoost model is performing best on the train set as well as test set with 88.7% accuracy. Therefore, the author will be using XGBoost to predict our values. With high accuracy score, it could better retrieve the authentic credit status of customers, which allows banking commercials to avoid the risk of bankrupting and simultaneously guarantee the benefits.

Outcomes

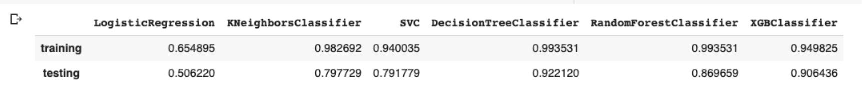
a). We checked the overall accuracy of the six classifiers with total population and sub-populations respectively. For overall accuracy, we generated the accuracy matrix shown below. The matrix for overall accuracy level reveals that for training set accuracy, DecisionTreeClassifier and RandomForestClassifier have the highest accuracy value (0.994353) while LogisticRegression has the worst performance (0.646520). For the test set, the XGBClassifier has the best performance (0.887664) while the LogisticRegression has the worst performance (0.508532). We will stick to the result from the test set since training set accuracy may not be reliable.



	LogisticRegression	KNeighborsClassifier	SVC	DecisionTreeClassifier	RandomForestClassifier	XGBClassifier
training	0.646520	0.978404	0.934905	0.994353	0.994353	0.956426
testing	0.508532	0.716139	0.750267	0.833274	0.779950	0.887664

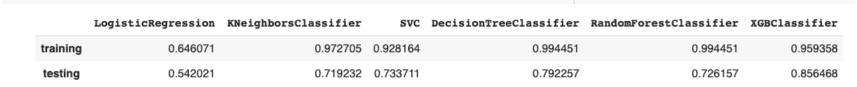
overall accuracy

As to the sub-population accuracy, we investigated the sub-population accuracy for male and female groups. Generally speaking, the accuracy for male group is much higher than female group, which might resulted from the disparity impact of the classifiers: the classifier mistakenly identifies a fair proportion of females as bad credit customers while they are actually good credit customers who may never default on loans. The accuracy level for male group ranges from 0.506220 (LogisticRegression) to 0.922120 (DecisionTreeClassifier) while the accuracy for female group ranges from 0.542021 (LogisticRegression) to 0.856468 (XGBClassifier).



	LogisticRegression	KNeighborsClassifier	SVC	DecisionTreeClassifier	RandomForestClassifier	XGBClassifier
training	0.654895	0.982692	0.940035	0.993531	0.993531	0.949825
testing	0.506220	0.797729	0.791779	0.922120	0.869659	0.906436

male accuracy

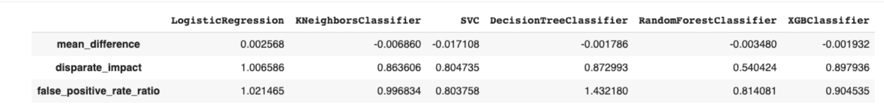


	LogisticRegression	KNeighborsClassifier	SVC	DecisionTreeClassifier	RandomForestClassifier	XGBClassifier
training	0.646071	0.972705	0.928164	0.994451	0.994451	0.959358
testing	0.542021	0.719232	0.733711	0.792257	0.726157	0.856468

female accuracy

Moreover, for different classifiers, we could conclude from the statistics that LogisticRegression has the least accuracy level for all populations (total population and sub-population). The XGBClassifier has the highest accuracy level for all test groups in overall, female and male populations. Therefore, the ADS system may employ the XGBClassifier as the basis of its algorithm. Although DecisionTreeClassifier and RandomForestClassifier have highest score for overall training set, we decline their importance as our main classifier since their performance on test sets are not satisfying.

b). In terms of fairness and disparity matrices, the models are evaluated by mean difference value, disparate impact and false positive ratio between different sub-populations, specifically speaking, male and female sub-population. We firstly set the male group as the privileged group while the female group is labeled as unprivileged. For mean difference value, we could observe that most of the the mean difference values are negative regardless of the classifier. This indicates that the disparate impact exists and all of the classifiers generates less favorable outcomes for female group. Similarly, the disparate impact value for most of the classifiers are smaller than one, meaning the classifiers are in favor of the privileged group (male group) and generate favorable outcome for this demographic group. Therefore, we could get the conclusion that there is apparent disparate impact and bias between male and female groups and the ADS system obviously favors the male group as the privilege group. For false positive rate ratio, most of the classifiers shows a ratio smaller than one, meaning that the false positive rate for male group is much larger than female group. As a result, a larger proportion of male group than female group is labeled as good credit while they are very likely to default on loans. The statistical disparity on false positive rate ratio indicates that females are unprivileged and biased.

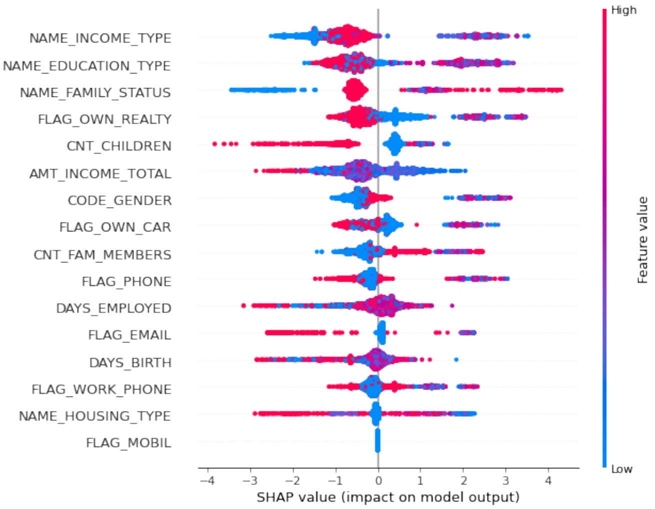


	LogisticRegression	KNeighborsClassifier	SVC	DecisionTreeClassifier	RandomForestClassifier	XGBClassifier
mean_difference	0.002568	-0.006860	-0.017108	-0.001786	-0.003480	-0.001932
disparate_impact	1.006586	0.863606	0.804735	0.872993	0.540424	0.897936
false_positive_rate_ratio	1.021465	0.996834	0.803758	1.432180	0.814081	0.904535

disparity matrix

The matrices that we selected are indicator showing the disparity impact on different demographic groups: disparate impact and mean difference value give lights on whether the classifier assigns favorable outcomes for privileged groups and assign unfavorable outcomes for unprivileged groups. False positive rate ratio is an indicator showing the extent of statistical disparity: a value less than one means that unprivileged groups are biased against and underestimated for favorable outcomes. Thus the statistical disparity is implied if the value falls far away from one.

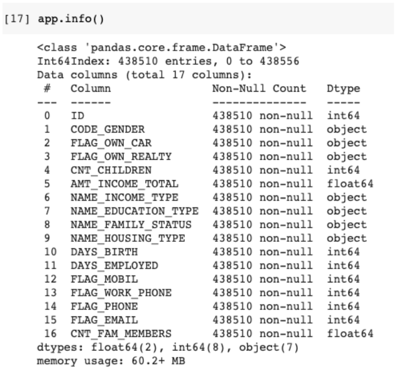
c). Given that the author applied the XGBclassifier, which has the highest accuracy score, to predict the credit status, we then employed SHAP explainer and LIME explainer to inspect the principles behind such algorithm.



shap summary

From the shap summary plot above, we could get a general idea of how different features affect the evaluation of credit status. For example, in terms of ’name income type’, there is a lump of red points to the left of the 0 shap value line, indicating that it is more likely to be classified as good credit when the customer has a high income type value. In contrast, blue points have a more uniform distribution over the shap value, which indicates that there might be some randomness in the contribution under the lower income type value.

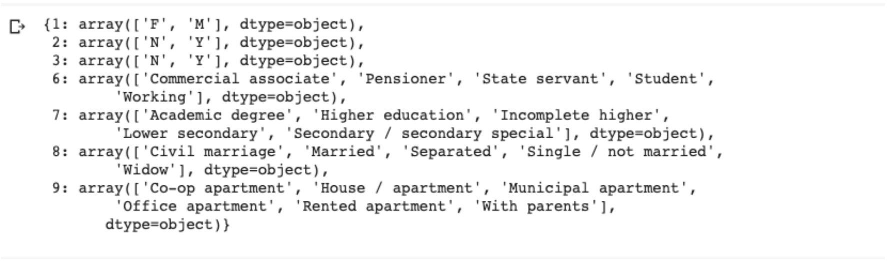
As for the meaning of income type value, we could learn it from the encoder package, which assign ascending values to category list of each feature (shown in the table below). So we could then see which category that the feature value corresponds to. Basically, commercial associate and pensioner are considered to have low feature value, while state servants, students, and working are assigned with high feature value. Thus, state servants, students, or stable workers will be more likely to be classified as having good credit status. Furthermore, since the lower-income type value has both distributions in negative shap value and positive shap value, we could infer that a commercial associate might be classified as having good credit while a pensioner might have a disadvantage over the credit evaluation.



```
[17] app.info()

     <class 'pandas.core.frame.DataFrame'>
     Int64Index: 438510 entries, 0 to 438556
     Data columns (total 17 columns):
      #   Column               Non-Null Count   Dtype  
     ---  ------               --------------   -----  
      0   ID                   438510 non-null  int64  
      1   CODE_GENDER          438510 non-null  object 
      2   FLAG_OWN_CAR         438510 non-null  object 
      3   FLAG_OWN_REALTY      438510 non-null  object 
      4   CNT_CHILDREN         438510 non-null  int64  
      5   AMT_INCOME_TOTAL     438510 non-null  float64
      6   NAME_INCOME_TYPE     438510 non-null  object 
      7   NAME_EDUCATION_TYPE  438510 non-null  object 
      8   NAME_FAMILY_STATUS   438510 non-null  object 
      9   NAME_HOUSING_TYPE    438510 non-null  object 
      10  DAYS_BIRTH           438510 non-null  int64  
      11  DAYS_EMPLOYED        438510 non-null  int64  
      12  FLAG_MOBIL           438510 non-null  int64  
      13  FLAG_WORK_PHONE      438510 non-null  int64  
      14  FLAG_PHONE           438510 non-null  int64  
      15  FLAG_EMAIL           438510 non-null  int64  
      16  CNT_FAM_MEMBERS      438510 non-null  float64
     dtypes: float64(2), int64(8), object(7)
     memory usage: 60.2+ MB
```

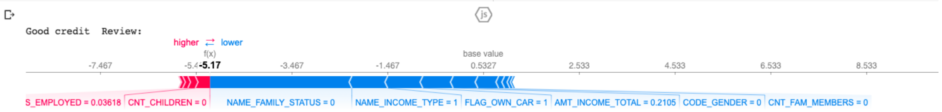
All the features



```
{1: array(['F', 'M'], dtype=object),
 2: array(['N', 'Y'], dtype=object),
 3: array(['N', 'Y'], dtype=object),
 6: array(['Commercial associate', 'Pensioner', 'State servant', 'Student',
        'Working'], dtype=object),
 7: array(['Academic degree', 'Higher education', 'Incomplete higher',
        'Lower secondary', 'Secondary / secondary special'], dtype=object),
 8: array(['Civil marriage', 'Married', 'Separated', 'Single / not married',
        'Widow'], dtype=object),
 9: array(['Co-op apartment', 'House / apartment', 'Municipal apartment',
        'Office apartment', 'Rented apartment', 'With parents'],
       dtype=object)}
```

information of categorical features

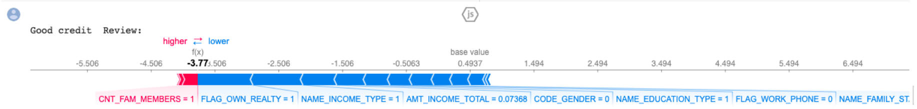
When we specify a customer in our prediction, say the 50th client in our test dataset, we then employed the force plot to analyze how specific feature values contribute to the classification. According to the plot below, we could see that the true status of this client is. The main features that contribute to good credit classification are family status, income type, and property status of the car. By applying the same methods as above, we could then have a general grasp of the principle behind the classifier.



```
Good credit  Review:
```

the 50th customer's view

We also found the misclassified customers and applied a force plot to one of them to see what features contribute to the classification. From the plot below, we could see that the credit status of this customer is actually bad. Still, the classifier classified such customer as having good credit since the overall shap value here is negative. Therefore, by looking at the main attributes to the good credit classification, we learn that the income type, family status, and total amount of income here mislead the classification, which pushes us to think about the relationship and influence between these features. As a result, we could further check the stability and the robustness of such a classifier. According to the discussion above, we tend to think this classifier is unstable since many features actually have randomness in the contribution to the classification, which could generate bias or mistakes in the results. Thus, the XGBClassifier applied here to predict credit status may not be the best choice and needs further inspection and improvement.

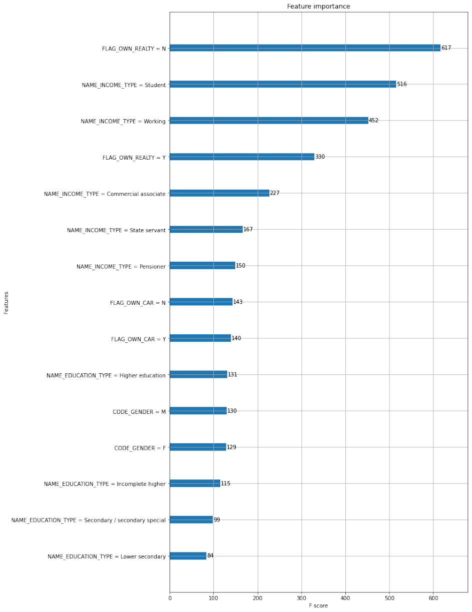


```
Good credit  Review:
```

the 36th customer's view

Finally, we inspected the ADS with LIME, which is an alternative way of explaining the XGBooster classifier. Specifically, we plot the importance of each feature based on fitted trees (xgbooster with 200 n_estimators and max depth of 5). From the plot below, we could learn that the status of property contributes most in classifying the client as having good credit, which corresponds to our common sense that if a customer does not have a property, then he or she might not be able to get the loan. The second significant feature is the income type. If a student would like to apply a loan, it is likely that he encounters the problem of affording the tuition. Since students are backed by families, and also have the most potential to improve themselves and earn the money in the future, therefore, they would be regarded as having the ability to pay back the loan, which supports the principle of classifying them as having good credit. The third important feature is working in the income type. With stable work, the client will have a higher possibility of being classified as having good credit. These analyses of income type correspond to the explanation from the shap summary polt, which further support our understanding of this classifier.

In conclusion, we regard this ADS as employing relatively appropriate principles to predict the credit status. Nevertheless, by generating different synthetic dataset to feed in the classifier and inspection, the explanation of each features are slightly different, which informs that the ADS might not be robust enough to be applied. Therefore, the ADS still needs further research and improvement before applied in the real world.



Lime importance plot

Summary

a). I think the data is not appropriate for ADS. The original dataset is highly oversampled in terms of the credit status, which will bring about pre-existing bias before fitting in the classifier. Moreover, there are numerous missing values in the 'OCCUPATION_TYPE' feature, and the author simply dropped all the inappropriate data, but it could be better addressed by further researching since there might be some apriori interpretation of the features here.

b). The model the author finally employed exhibited an accurate but somewhat biased and non-robust prediction of the dataset. Our initial analysis about this set of prediction models uses mean difference value, disparate impact and false positive ratio as three indicators to measure the overall bias level of the models. The fairness analysis found out that generally speaking, the prediction models the author employed shows a preference for the privileged group (male) while biasing against the unprivileged group (female). In terms of stakeholders benefiting from these measures, ADS hold the view that unprivileged group (female) and the bank would surely be benefactors. The unprivileged group would benefit in a way that the systematic bias from the prediction is brought to the attention of more people as it is clearly illustrated by our measurements. On the other hand, the bank could have less riskier clients when taking our measurement into consideration because they are aware of the bias in the model, which could otherwise undermine the risk management of commercial banking.

c). It is not comfortable to deploy this ADS in the public sector and this ADS system should be contained in the industry sector. Firstly, the ADS system is not immune to re-purposing. When employing this system, different subjects may have different intentions which is beyond our control. For example, some gamble companies may use this system to check if it is likely for a person to pay off the debt. Using this ADS system could increase the efficiency of the gambling company and more people could get addicted to gambling. Secondly, If we would like to deploy this ADS system in the public sector, we will have to enhance the transparency of the system. As a result, more information of the ADS system would be released. In this sense, we should consider the issue of data protection and ensure that there is no privacy compromising when the data is interpreted to the public. However, there is literally no data protection approach in the ADS system so far: no synthetic dataset, not covered sensitive feature, etc. In this way, the dataset and the model is vulnerable to data hackers that seek opportunities to break into the privacy of the data. Thirdly, as the vendor of the data, we could not trust the data users (analysts) in public sector. It would be easier to contain the risk of misusing the data and enforcing regulations in the industry sector than in the entire public sector. In the industry sector, the use of data and ADS system could be constrained to a small extent that minimizes data hacking and re-purposing of the ADS system.

d). Though there is little information about data collection in this case we picked, we hold still to the principle that all data collection process should be ethical: we only collect data from people who give their consent and these people are fully informed about the possible usage of their information. Transparency and informant are two basic principles that we must stick to when collecting data. In terms of data processing, the ADS system could improve fairness by mitigating the great disparity between different demographic groups (male and female in our case). The ADS could do reweighing process. It could set different threshold for different demographic groups: set lower threshold for positive outcome to female groups since they are biased against and set relatively higher threshold to male groups. In this way, more female could get the loan because the condition threshold for granting them a loan drops and the overall level of disparity could be mitigated to some extent. Moreover, the ADS system could use synthetic dataset and remove some features with high correlation to gender to mitigate the bias: for example, the system could turn all the gender of candidates to male and remove some relational features such as flag own car (male may be more likely to have a car than female).