Financial Bias and Predicting Loan Default

Intro

Financial services are using artificial intelligence (AI) and machine learning (ML) methods for a wide range of operations and customer journeys. These include mortgage lending, savings, insurance, fraud detection, etc. However, if the algorithms used to make the predictions for individuals' suitability for a specific service contains bias, it could impact negatively. Such as being from a specific race, gender or age should not cause any unfair treatment. Therefore, there are ethical and regulatory needs to prove that models used for these financial activities are fair and unbiased.

To provide a quick overview of ML and AI. Machine learning is a subset of AI that uses statistical models to draw insights and make predictions. Using structured or unstructured data, it allows trends or relations to be calculated allowing for conclusive decision making. In the finance industry big data is a huge asset in the AI process. In order to use ML methods to optimize data and find new relations or to understand a specific client better, big data brings the edge of large-scale information not usually present in existing financial models. However, again this data or the algorithm being used might contain bias which can drive bias into the final prediction, such as whether a person is eligible for a loan or not.

Problem Outline

To understand how data can drive bias and to demonstrate a simple way to mitigate it, our Credit Card Loan default dataset was used. The aim was to predict if an individual would default(1) or non-default(0) on their loan based on the information provided by the applicant. Some of the features included were gender, occupation type, family size, etc. The approach towards the model building for the dataset was divided into three sections:

1) Pre-processing: To conduct an Exploratory Data Analysis (EDA) which would summarise the main characteristics in the dataset. This is done to see what the data tells us beyond the modelling and hypothesis test. Such as any imbalance in the dataset; missing values; quick visual aid to see any straightforward relationships between other features and the target variable.

2) In-processing: To conduct a simple statistical bias test to see the disparity between categorical variables such as the difference for gender to default on a loan according to the dataset. Usually a high disparity of about 20% or more between the labels would demonstrate bias.

3) Post-processing: To model the dataset and finalise features using PyCaret, an ML library to obtain predictions. This will allow the gender predictions to be compared and allow any further feature engineering if there is a possibility to mitigate any bias present.

The focus was to try and mitigate any gender bias when building a predictive model.

Theory

In order to achieve fair machine learning, we first need a mathematical description of bias. We can obtain such a description with a consideration of the three following terms: (1) anti-classification – model training doesn't include gender as a feature, or proxies thereof; (2) classification parity – model evaluation metrics significant to the problem are equal across male and female subgroups; and (3) calibration – the number of individuals with a given risk score is proportional across male and female subgroups.

We can define each of these mathematically if we first define the following:

- $x_i \in \mathbb{R}^n$ – where x_i is the features of sample i.
- $x = (x_g, x_a)$ – where x is the concatenation of gender, x_g and all other features, x_a. Note that x_g may also include any other feature deemed to be a gender proxy.
- $y_i \in [0, 1]$ – where y_i is the target variable for sample i, for example, $y_i = 1$ represents an individual who defaulted on their loan etc.
- $d : x_i \to y_i$ – where d is a map from the feature space to the binary target variable.
- $s(x_i)$ – risk score produced by the model such as the probability of default given observable features x_i.

With these in mind, we can formulate anti-classification as:

$$d(x) = d(x') \quad \forall \quad x, x' \text{ s.t. } x_a = x'_a.$$

This means all samples with the same features excluding gender and gender proxies will be subject to the same final decision. For classification parity, we can consider any metric or combination of metrics that can be derived from a confusion matrix once a model has been trained. Given the nature of our problem, we assume the action of approving a loan to someone who later defaults to be the costliest action, meaning we want to minimise the number of false negatives. One appropriate metric would be recall – the ratio of true positives to the sum of true positives and false negatives. In which case, classification parity of false negatives can be expressed as:

$$P(d(x) = 0 \mid y = 1, x_g) = P(d(x) = 0 \mid y = 1).$$

This means false negative rates should be the same for male and female predictions. Finally, calibration may be defined as

$$P(y = 1 \mid s(x), x_g) = P(y = 1 \mid s(x)).$$

This means the proportion of individuals with a given risk score who actually defaulted on their loan is the same across male and female subgroups.

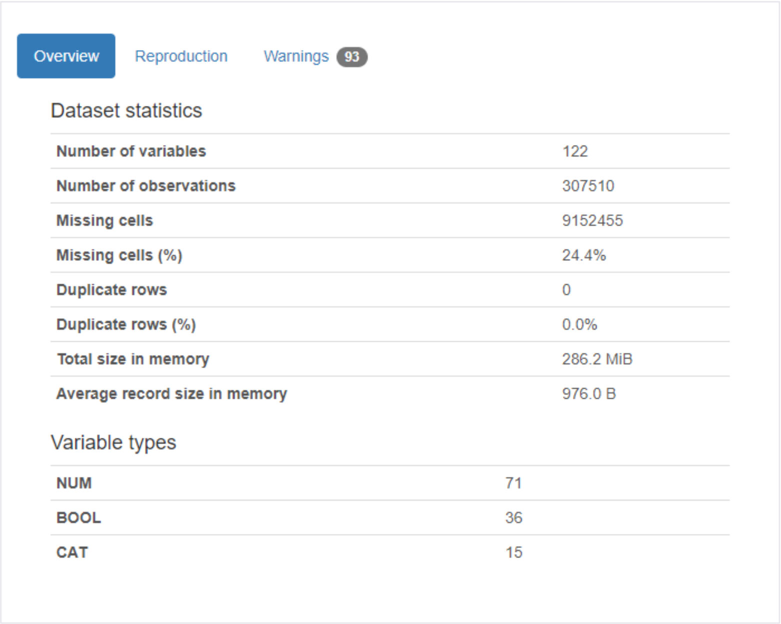
In building a model that satisfies these three definitions, we aim to mitigate bias when it comes to data-driven decision making with this simple approach. We do recognise that alternative practices exist and are aware of the limitations that also exist with this proposed framework. We do not propose this as a best practice, but merely exploring the simplicity and effectiveness of this approach.

Pre-Processing

The dataset consists of 300,000+ observations with 100+ variables. Some of these variables can be categorised into the following:

- Demographic data: gender, marital status, income, level of education, employment history etc.
- Other personal data: housing information, population of region where the client lives, rating of the region, if their home address is the same as their work address etc.
- Information about the loan: contract type, loan annuity, credit amount of the loan etc.
- Anonymised data: some of the data has been anonymised and includes information from external data sources.

Pandas profiling can give us a quick overview of the data including information on missing values, data types and variable distributions:



Dataset statistics

Number of variables	122
Number of observations	307510
Missing cells	9152455
Missing cells (%)	24.4%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	286.2 MiB
Average record size in memory	976.0 B

Variable types

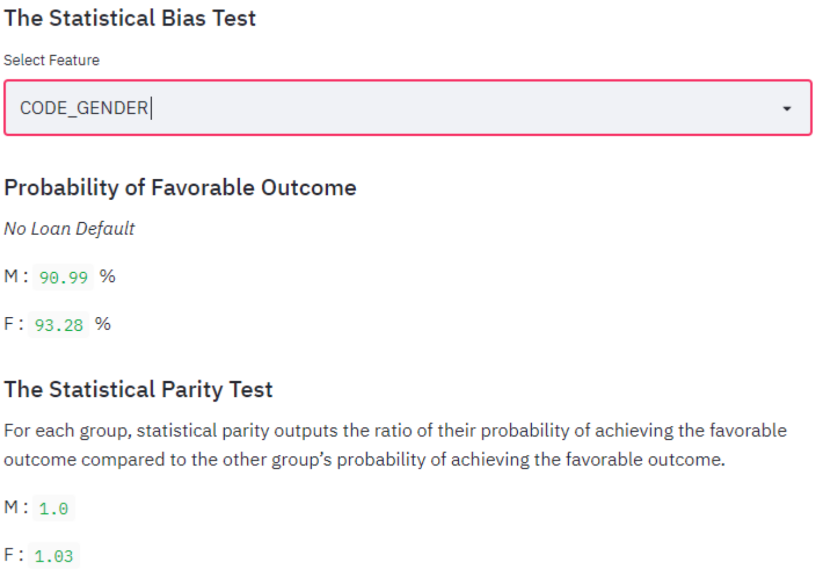
NUM	71
BOOL	36
CAT	15

There are many missing values in the dataset, with 8 of the 15 categorical (non-binary) variables possessing missing data. For simplicity, missing categorical data was imputed with a constant ‘None’ placeholder. For numerical variables, multivariate imputation by chained equations (MICE) was used to impute missing data using Sklearn’s iterative imputer class. MICE uses the feature space to model features with missing values as a function of the other features. It is a more statistically sound way to impute missing data as opposed to standard mean/median imputation but only works with numerical data.

The profile also shows that the target distribution is severely imbalanced with 92% of the labels being No Default(0). Under sampling was used to balance the training data with Imblearns’ random under sampler class (SMOTE and over sampling provided poor recall scores for this problem).

In-Processing

To test for sample bias, firstly the probability of favourable outcome was calculated for the gender feature. It showed that 93% of the females in the dataset did not default on their loan compared to 91% men. Even though the difference was small this could drive bias during the final predictions. The model can learn that solely based on the fact that a person is male he is more likely to default giving an unfair advantage. A statistical parity test was then done to obtain a ratio comparing the output for the favourable outcome for both the labels. So a label which in this case is female with a ratio of more than 1 would indicate that it is more likely to obtain the outcome. This is indeed the case as shown below.



The Statistical Bias Test

Select Feature

CODE_GENDER

Probability of Favorable Outcome

No Loan Default

M : `90.99` %

F : `93.28` %

The Statistical Parity Test

For each group, statistical parity outputs the ratio of their probability of achieving the favorable outcome compared to the other group’s probability of achieving the favorable outcome.

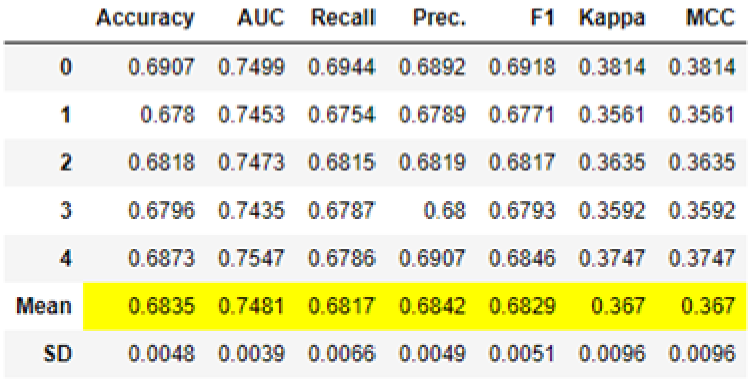
M : `1.0`

F : `1.03`

Post Processing

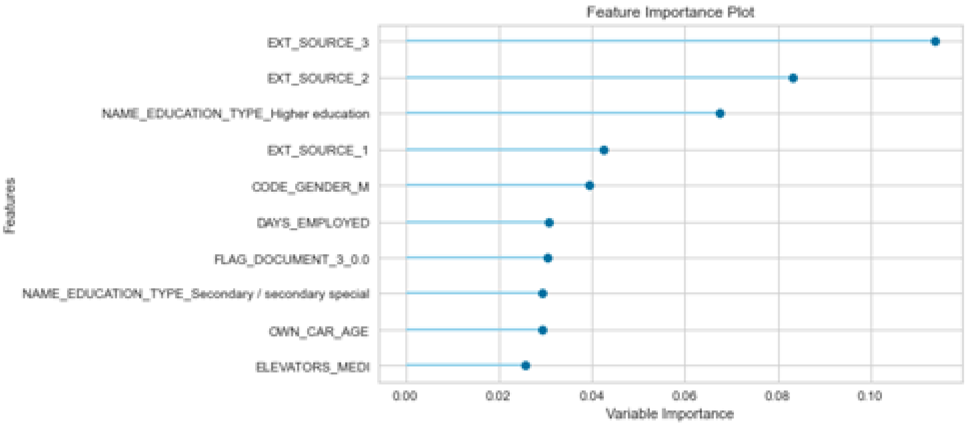
Final pre-processing tasks were carried out in the PyCaret environment. PyCaret is an open-source machine learning library that allows us to quickly experiment with different machine learning pipelines to find an optimal model with little coding. The final pipeline standardized the dataset before a final round of feature selection using feature importance from Random Forest and Adaboost algorithms. Categorical features were also one-hot encoded, and a 70-30 train-test split was used.

An XGBoost classifier was fitted to the training data with 5-fold cross-validation. The performance of each fold is displayed below.



	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.6907	0.7499	0.6944	0.6892	0.6918	0.3814	0.3814
1	0.678	0.7453	0.6754	0.6789	0.6771	0.3561	0.3561
2	0.6818	0.7473	0.6815	0.6819	0.6817	0.3635	0.3635
3	0.6796	0.7435	0.6787	0.68	0.6793	0.3592	0.3592
4	0.6873	0.7547	0.6786	0.6907	0.6846	0.3747	0.3747
Mean	0.6835	0.7481	0.6817	0.6842	0.6829	0.367	0.367
SD	0.0048	0.0039	0.0066	0.0049	0.0051	0.0096	0.0096

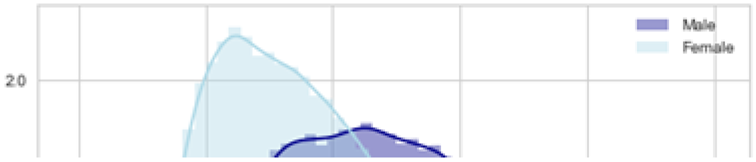
The model achieved a mean training accuracy and recall of approximately 68% and 0.68, respectively. The small standard deviations are also a good sign – the training scores have converged. We can also look at feature importance – how valuable each feature was relative to other features in the model training, put simply.

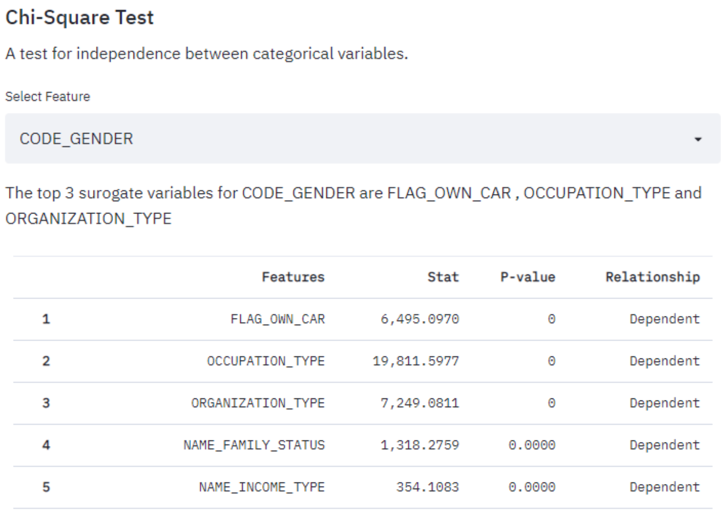


We can see the external source data proving to be most useful, as well as the level of education one has received. The appearance of gender, however, presents a concern as this shows the model has learnt the gender of the applicant to be important. We should then expect the distribution of risk scores to be different across male and female subgroups when we evaluate our model on the test set, as shown below.

This does indeed appear to be the case, where the risk score is the probability of default as predicted by the model. We can see the distribution for the female subgroup has a more defined peak at a lower risk score in comparison to the male distribution. One could employ gender-specific risk thresholds to satisfy calibration, but given the probability distributions here, the difference in these thresholds would not be equitable. A quick calculation also shows that classification parity is violated with male and female recall scores at 0.77 and 0.63, respectively. We may then conclude that our model violates all three of our bias definitions.

The first step in going about fixing this is to first satisfy anti-classification, where we remove gender and gender proxies from the equation. It is important to find such proxies because we can incorporate gender bias into our model if certain features exhibit gender differentials. To find gender proxies we can use a chi-test where we treat gender as a target variable. Results of the chi-test are displayed below, which shows the chi-statistic and the corresponding p-value.





Chi-Square Test

A test for independence between categorical variables.

Select Feature

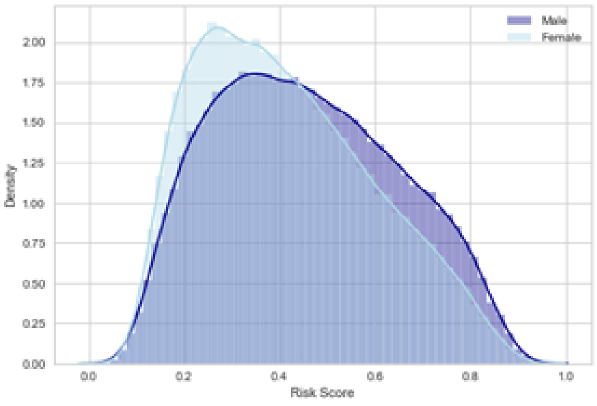
CODE_GENDER

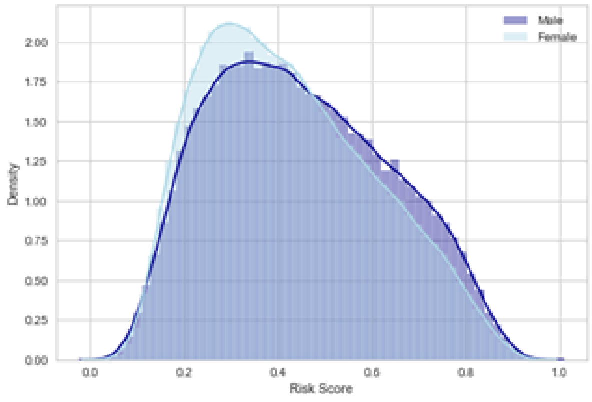
The top 3 surogate variables for CODE_GENDER are FLAG_OWN_CAR , OCCUPATION_TYPE and ORGANIZATION_TYPE

	Features	Stat	P-value	Relationship
1	FLAG_OWN_CAR	6,495.0970	0	Dependent
2	OCCUPATION_TYPE	19,811.5977	0	Dependent
3	ORGANIZATION_TYPE	7,249.0811	0	Dependent
4	NAME_FAMILY_STATUS	1,318.2759	0.0000	Dependent
5	NAME_INCOME_TYPE	354.1083	0.0000	Dependent

Based on the magnitude of the statistic, this test shows the features most correlated with gender in this dataset are occupation and organization type, whether the applicant owns their own car or not and hence the age of the car.

Below then shows the distribution of risk scores across the gender subgroups when we train a model without gender and when removing proxies also.





We can see straight away that distributions across the subgroups are now more equitable, with the no gender and proxy scenario providing the most equitable outcome. If one were to employ gender-specific thresholds now, the thresholds would be much closer together and hence calibration can be more easily satisfied. When it comes to classification parity, our no gender only model obtained male and female recall scores of 0.73 and 0.66, respectively, compared to the no gender and proxy model which obtained a more equitable result of 0.70 and 0.67, respectively.

Conclusion

Bayezian has then shown with this simple approach to mitigating bias, we have obtained a more equitable model without significantly impacting overall performance metrics. Such metrics can be expected to decrease, however, since we are removing important variables from the feature space. Hence, the obvious drawback with this method is the trade-off with model performance. To further assess the effectiveness of this approach, other statistical limitations should be investigated. Nonetheless, this is a simple and somewhat effective approach which can be easily implemented and can be seen as a starting point to mitigating bias in data-driven decision making.