Group DL2

Big Data in Finance - Submission 1

08.06.2020

Tackling Discrimination in Machine Learning Algorithms

Problem

In a world in which every human being is supposed to be free and have equal opportunities, discrimination keeps tearing our society apart. Based on prejudices and biases, people tend to treat each other differently, often in a worse manner, if they belong to a different group or present different traits, such as skin color and sex.

This discrimination is causing further social and economic inequality and this negative loop keeps reinforcing itself within every society. Bias and prejudices are rooted in our decisions as well the data originated from them. Historically vulnerable groups, like women and minorities, have been treated differently since the beginning of times. We made some steps forward but yet this is not enough. We can't afford to let these discriminatory biases spill over into our digital future. If we teach our machine learning algorithms as we taught our society in the past millennia, we'll never be able to fix this **injustice.**

As we enter an automated world, where we rely on machines to carry out decisions instead on ourselves, we must watch out whether our biases will leak into the algorithms. Research by the Journal of Big Data clearly states that "Big Data analytics such as credit scoring and predictive analytics offer numerous opportunities but also raise considerable concerns, among which the most pressing is the risk of discrimination."[1]

[1] Favaretto, M., De Clercq, E. & Elger, B.S. Big Data and discrimination: perils, promises, and solutions. A systematic review. *J Big Data* 6, 12 (2019). https://doi.org/10.1186/s40537-019-0177-4

Objective

Our objective is to fight data-driven inequality and discrimination which are defining social life in the algorithmic age.[2] Our analysis focuses on eliminating, or whether not possible at least minimizing, the effect of bias on human decisions. In this context, we aim at fostering a fair Machine Learning Model that does not take into consideration the bias on gender or other types of minorities. Through our data exploratory analysis, we aim at understanding whether the bias is present and derive the causes and consequences of discrimination. Our objective is threefold: First, identify barriers to fair data-mining. Second, detect and minimize wrongful discrimination, and lastly, investigate possible solutions to this rising problem.

Dataset & Methodology

We obtained a loan dataset from Kaggle that was made for pedagogic purposes.
"It emulates a common loan default prediction task, but the data is generated in such a way that default prediction machine learning models are likely to be biased against women and minorities." [3] The data consists of 15 personal attributes that are usually evaluated to provide loans, implicitly included are two dummy variables to check whether the loan seeker is a woman or belongs to a minority in order to better evaluate if these traits affect the output. If so, the algorithm must be fixed to take into consideration these negative human biases. After conducting some initial exploratory analysis and data preparation, we will first use deep learning to predict the default of customers. Afterwards, we will evaluate the algorithm to assess whether it discriminates unfairly against women and minorities. This should be the case and we will try to subsequently alter the data and approach in order to minimize the bias based on gender and minorities and achieve a fair outcome while maintaining high prediction accuracy.

[2] Cornell Center for Social Sciences, "Algorithms, Big Data, and Inequality (2018-2021), Cornell University, 208

[3] Jannes Klaas, "Loan Default Model Trap", Kaggle, 2020

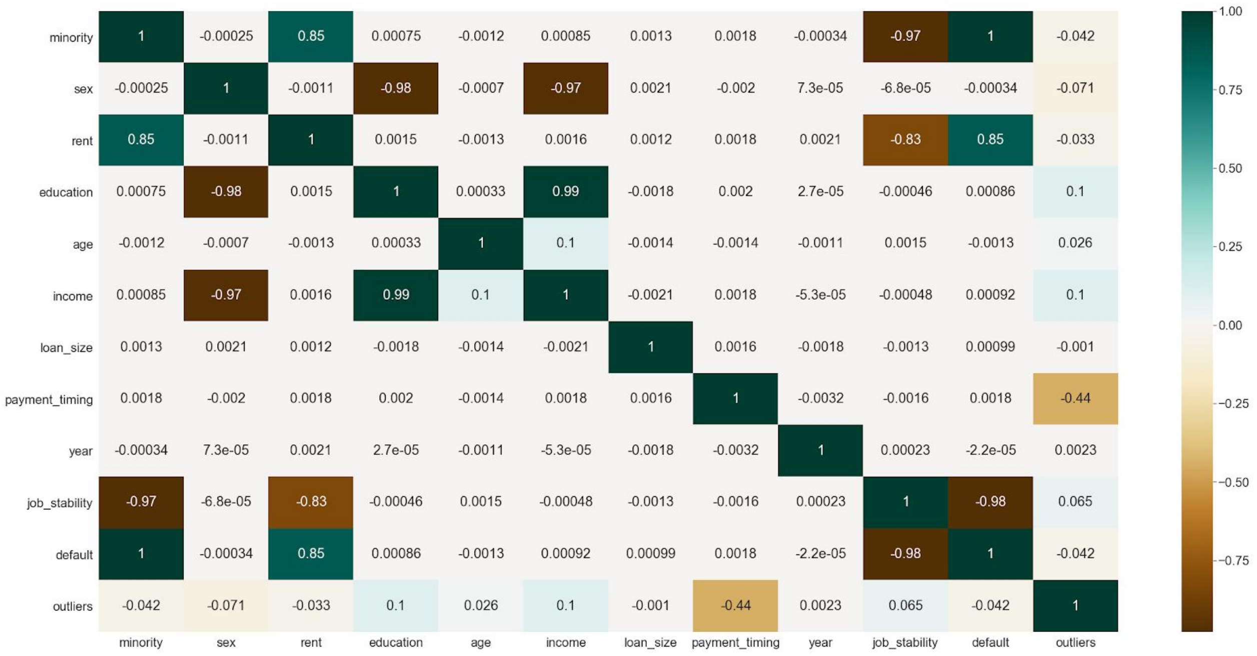
Results from exploratory analysis

The result of the exploratory analysis of the given dataset was threefold: First, we clearly identified that the data is potentially biased. Secondly, the dataset is not imbalanced. Lastly, defaults for gender are not unevenly distributed whereas for minorities they clearly are. The detailed code is documented in the attached Jupyter Notebook file. In the following the conducted steps will be briefly outlined.

Fortunately, the dataset did not require a lot of cleaning as there are no null values in the dataset. Subsequently, we tested whether there are any outliers in the dataset. We used the z-score technique and assigned 3 standard deviations as the threshold for identifying a value as an outlier. Only the column “payment_timing” has ~8,800 outliers, which we removed from the dataset.

Having ensured that our data does not include any factors that would impair our analysis, we focused on the exploration of its features. First, we checked if the dataset is imbalanced and found that it is not. The defaults are more or less evenly distributed. The same analysis was conducted focusing on the features gender and minority. While gender is also evenly distributed, the minority feature is clearly unevenly distributed. Almost all defaults are by minority clients.

In order to dive deeper into the correlations between our features, we applied a correlation matrix using seaborn (see full screenshot below). Obviously, correlation does not necessarily indicate causality but can provide a direction for our analysis. Some features are highly positively or negatively correlated. For example, being a minority is almost perfectly correlated with defaulting. Rent is also very closely correlated with defaulting and minorities. Furthermore, job security is negatively correlated with defaulting and with minorities. Not surprisingly, income and education are closely, positively correlated. Additionally, gender is negatively correlated with income and education.



	minority	sex	rent	education	age	income	loan_size	payment_timing	year	job_stability	default	outliers
minority	1	-0.00025	0.85	0.00075	-0.0012	0.00085	0.0013	0.0018	-0.00034	-0.97	1	-0.042
sex	-0.00025	1	-0.0011	-0.98	-0.0007	-0.97	0.0021	-0.002	7.3e-05	-6.8e-05	-0.00034	-0.071
rent	0.85	-0.0011	1	0.0015	-0.0013	0.0016	0.0012	0.0018	0.0021	-0.83	0.85	-0.033
education	0.00075	-0.98	0.0015	1	0.00033	0.99	-0.0018	0.002	2.7e-05	-0.00046	0.00086	0.1
age	-0.0012	-0.0007	-0.0013	0.00033	1	0.1	-0.0014	-0.0014	-0.0011	0.0015	-0.0013	0.026
income	0.00085	-0.97	0.0016	0.99	0.1	1	-0.0021	0.0018	-5.3e-05	-0.00048	0.00092	0.1
loan_size	0.0013	0.0021	0.0012	-0.0018	-0.0014	-0.0021	1	0.0016	-0.0018	-0.0013	0.00099	-0.001
payment_timing	0.0018	-0.002	0.0018	0.002	-0.0014	0.0018	0.0016	1	-0.0032	-0.0016	0.0018	-0.44
year	-0.00034	7.3e-05	0.0021	2.7e-05	-0.0011	-5.3e-05	-0.0018	-0.0032	1	0.00023	-2.2e-05	0.0023
job_stability	-0.97	-6.8e-05	-0.83	-0.00046	0.0015	-0.00048	-0.0013	-0.0016	0.00023	1	-0.98	0.065
default	1	-0.00034	0.85	0.00086	-0.0013	0.00092	0.00099	0.0018	-2.2e-05	-0.98	1	-0.042
outliers	-0.042	-0.071	-0.033	0.1	0.026	0.1	-0.001	-0.44	0.0023	0.065	-0.042	1

Additionally, we applied three different methods to get an initial idea on which features are important for predicting our outcome variable. Using the univariate chi, recursive elimination and tree based approaches, we were to extract some insights on feature importance. In all three models, minority is the most important feature. This is not surprising as we have already seen that it is perfectly correlated with default. Moreover, the ZIP codes, rent and job stability seem to be important for determining whether a client defaults or not.

In the next step, we will select the most appropriate features given their correlation and feature importance and identify the most promising variables for creating a deep learning model that accurately classifies customers as defaulting or not. Subsequently, we will analyse whether the model is biased against minorities and gender. Specifically, we will analyze the bias using four different approaches[4]:

1. **Demographic Parity:** Does our model positively classify females and minorities at the same rate as the total population?

[4] https://towardsdatascience.com/identifying-and-correcting-label-bias-in-machine-learning-ed177d30349e

2. **Disparate impact:** Same question as in previous approach but this time the model does not know which are the subgroups we are trying to assess and which data points belong to these groups.
3. **Equal opportunity:** Does the model have an true positive rate that is equal when comparing the total population to the subgroups investigated?
4. **Equalized odds:** Does the model have an true positive and false positive rate that is equal when comparing the total population to the subgroups investigated?

In the final step, assuming we identify a bias towards women and minorities in the dataset, we will correct the model to minimize the bias. Lastly, we will compare the performance and bias of the initial, raw model with the adjusted model. We aim to achieve a similarly high accuracy while minimizing the unjust bias in the model.

In a world that is biased by nature, tackling discrimination spillovers in machine learning will not constitute an easy task, however we do believe that "the journey of a thousand miles begins with one step" (Chinese Proverb). Afterall, we must not be satisfied by only doing things right, but rather by doing the right things.

APPENDIX

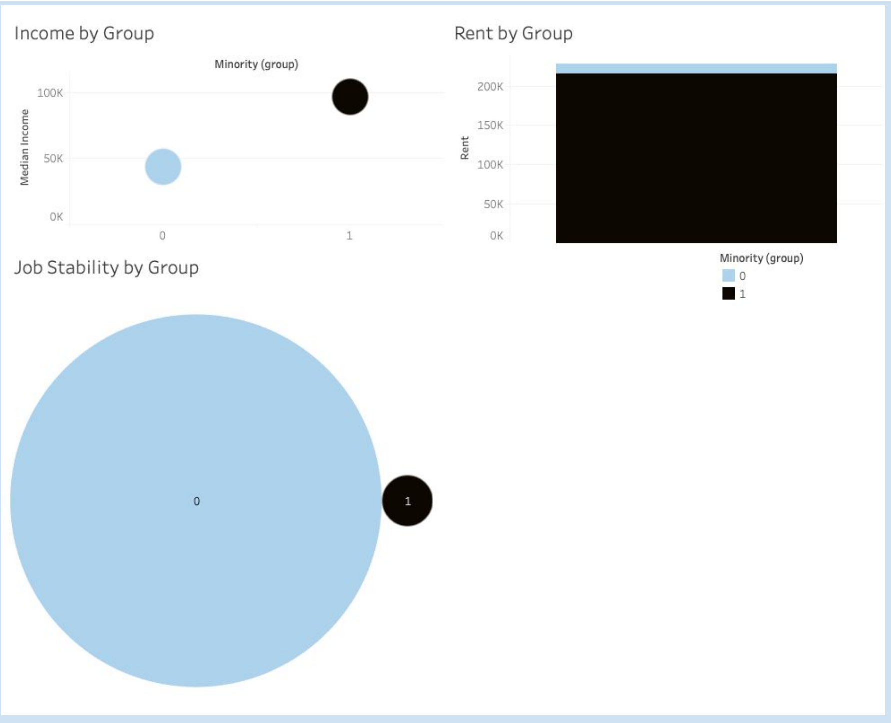


Figure 1: Exploratory Analysis by Group. We notice that minority groups tend to have less job stability than their counterparts.

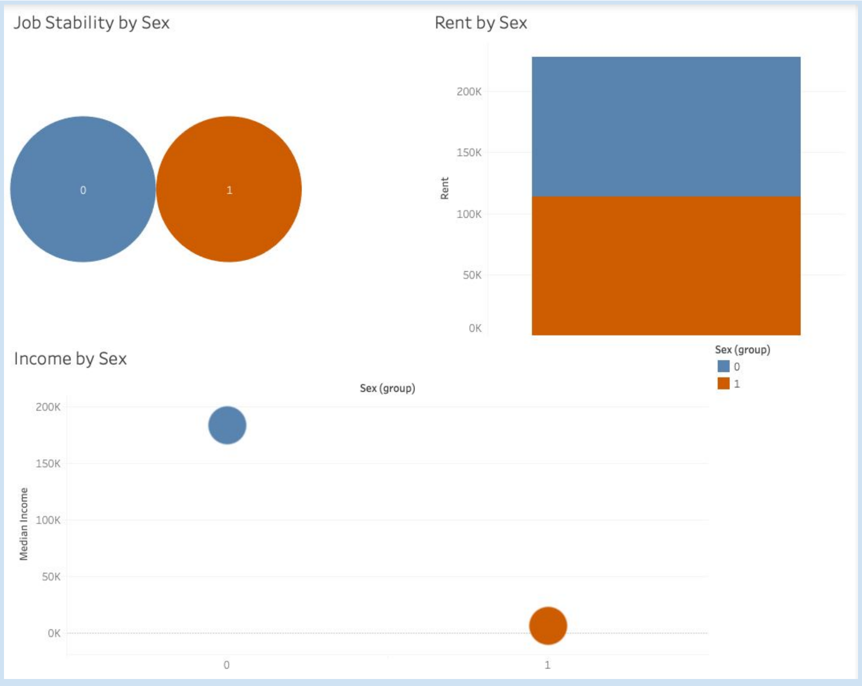


Figure 2: Exploratory Analysis by Gender. We notice that women tend to have much lower income than men.

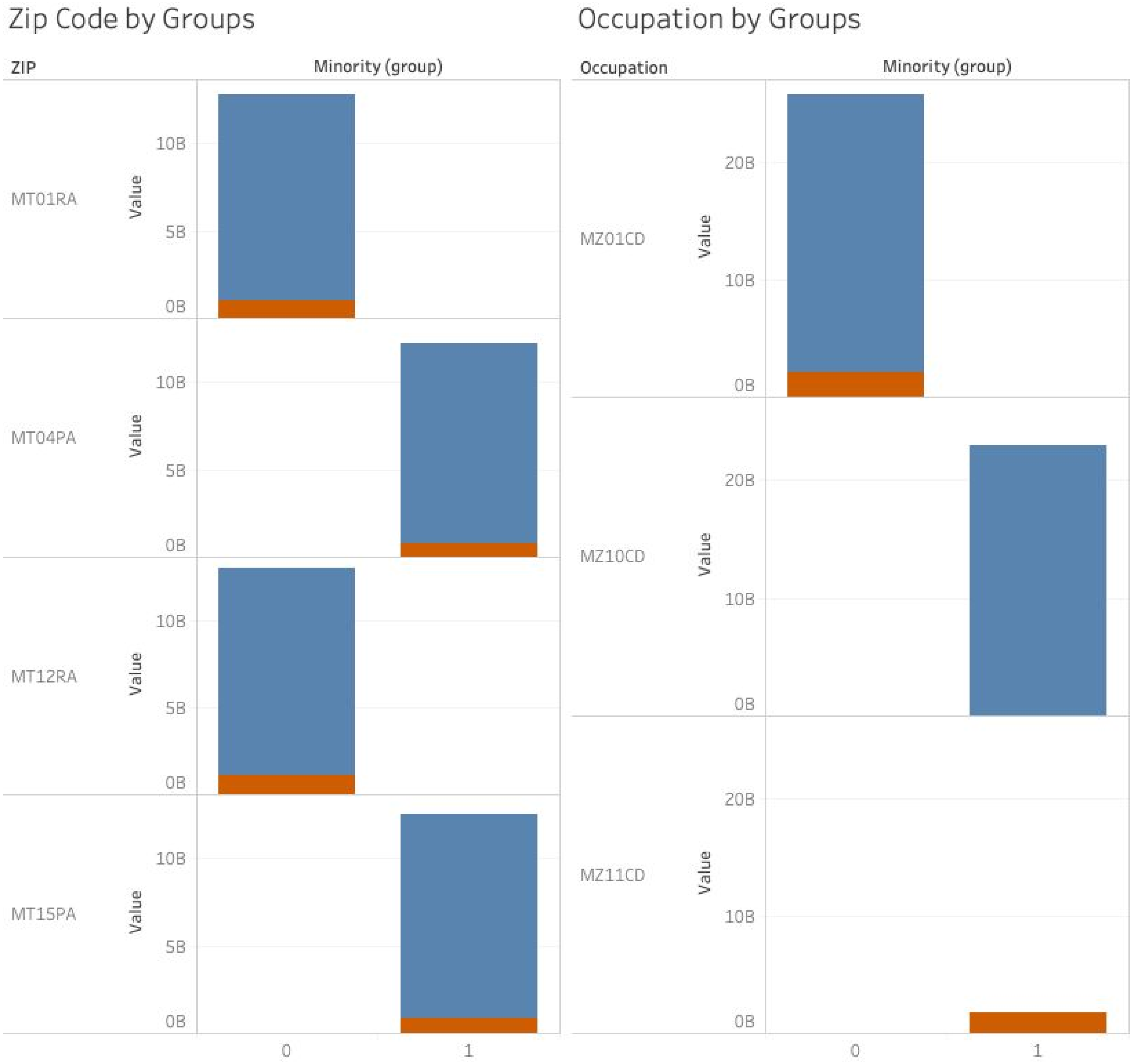


Figure 3: Exploratory Analysis by Groups and Gender. We notice that minority groups tend to have different jobs and live in different neighbourhoods, as indicated by the ZIP codes of each customer.