Date: 12/12/2019

Professor: Indrė Žliobaitė

The Ex-Presidents bank's program for exclusive customer

Tommaso Carraro, Alberto Gallinaro

Helsinki December 12, 2019

UNIVERSITY OF HELSINKI
Department of Computer Science

Contents

1 Context

For the purposes of the project we imagined of being the R&D department of the Ex-Presidents bank. Such bank offers benefits to the most exclusive customers (customers with an income grater than 50K dollars per year).
The benefits provided are:

- free personal financial consultant;
- reduced interest rate on loans;
- free American Express Black (exclusive access to airline lounges, private events).

We utilize a ML model in order to assess whether a customer is worth being guaranteed the benefits or not, also, we want to be fair, so sensible variables like race or gender mustn't have any impact in the final decision. Only personal merits should be taken into account.

For this project we decided to utilize the Adult Income Dataset that can be found at: https://archive.ics.uci.edu/ml/datasets/Adult.

2 Dataset description

The purpose of the Adult Income Dataset is to predict if an individual has an income higher than 50K given some characteristics such as age, education, occupation, marital status, relationship, race, sex, capital gain, capital loss, hours of work per week and native country. This dataset defines a binary classification task and the possible values of the target variable are:

- 1: if the individual has an income higher then 50K;
- 0: if the individual has an income that is less than 50K.

So, it is possible to observe that the target variable is a polar variable and in this case having an income higher than 50K is better than having an income that is less than 50K. Since this dataset has a target variable that is polar, we can begin talking about discrimination in the historical data that compose it.

The protected attributes of this dataset are sex and race. Our project is mainly focused on measuring and getting fairness on the sex attribute, due to time limitations. As you have seen, the dataset contains many attributes and we used all of them to train our models, but only few of them are used in our observations about fairness and discrimination. The attributes we used to asses fairness are:

- sex: as we already seen this is the protected attribute which our experiments are based on;

- hours of work per week: we used this attribute to see if males and females have similar merits in terms of quantity of work;
- years of education: we used this attribute to see if males and females have similar merits in terms of quantity of education;
- education level: we used this attribute to explore the biases more in detail, especially to see if there are some education levels (e.g. doctorate, high school, etc.) that are subjected to higher biases;
- occupation type: we used this attribute to explore the biases more in detail, especially to see if there are some occupation types (e.g. manager, sales, etc.) that are subjected to higher biases.

2.1 Minimal pre-processing

We performed a minimal pre-processing on our dataset:

- we removed all the rows which contained ambiguous values (e.g. ?);
- we removed two races ("Other" and "Amer-Indian-Eskimo") because they contained only few examples and especially because the bias we wanted to study was between black and white people.

3 Unfair dataset

3.1 Data exploration on raw dataset

We began our data exploration analysis with the plot in Figure 1.

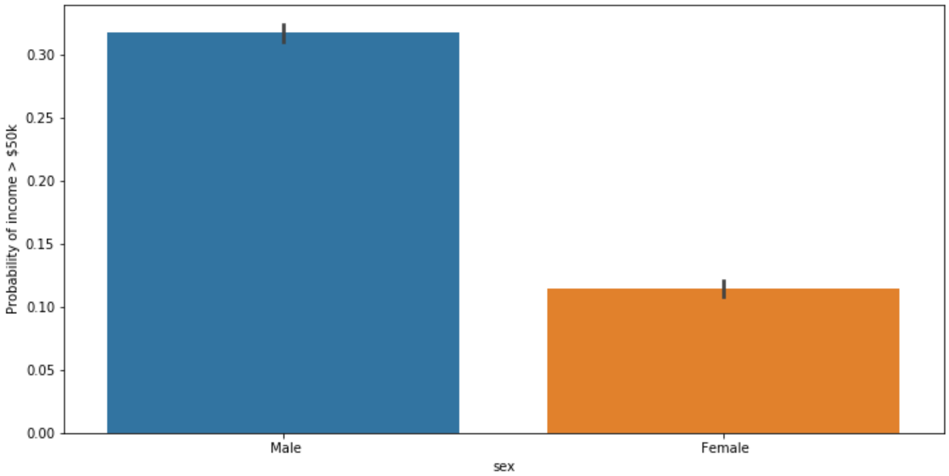

Figure 1: Probability of income higher than 50K for males and females in the raw training set.

It is possible to observe that there is an high bias towards males to have an income higher than 50K. This could be due to:

- explainable discrimination: females could work less hours during the week or could have less years of education compared to males, so they could have a smaller salary and then a smaller probability to have an income higher than 50K;
- bad discrimination: even if males and females have the same merits, it could be that males have higher salaries and then an higher probability to have an income higher than 50K. This could be due to the presence of bad discrimination on historical data that compose the dataset.

So, the aim of this project is to see if this bias is due to explainable discrimination, bad discrimination or a combination of the two. Furthermore, our goal is to implement some fairness aware strategies in order to remove the bad discrimination and make the dataset fair before training a predictive model on it. In fact, if a model is trained on a unfair dataset, it will provide unfair predictions as well.
After we have identified that bias on the sex attribute we plotted the graphs in Figure 2 and Figure 3 to see if in average males and females have the same merits in the dataset.

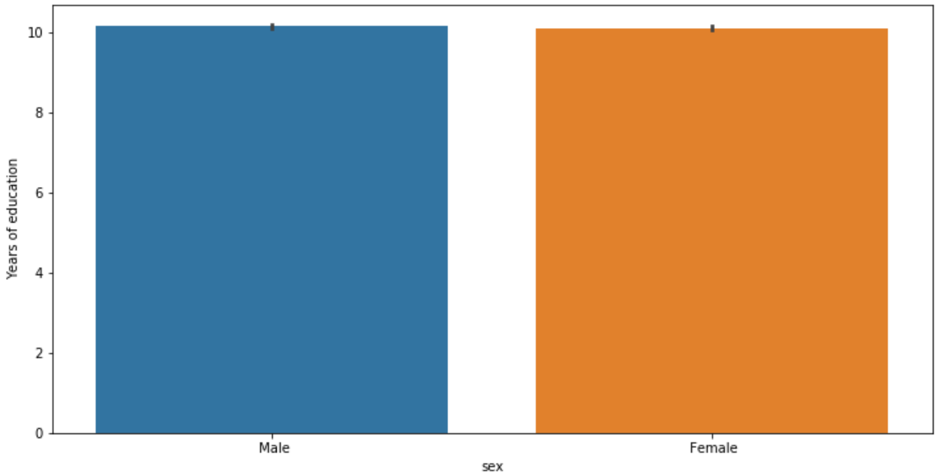

Figure 2: Average of years of education for males and females in the raw training set.

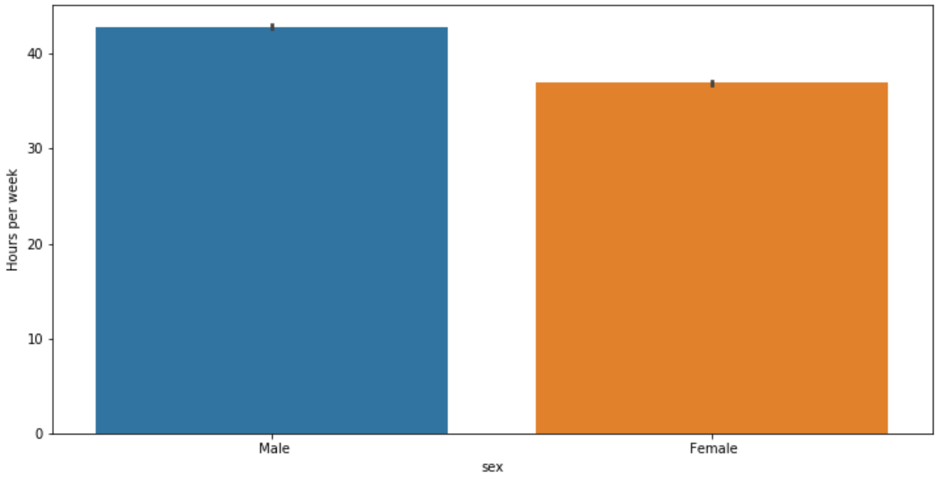

Figure 3: Average of hours of work per week for males and females in the raw training set.

From the graph in Figure 2 is possible to see that in average males and females have exactly the same merits in terms of years of education, while in the graph in Figure 3 is possible to see that in average females work about five hours less than males and this means that males and females have similar merits in terms of quantity of work.
The slight difference in the hours of work per week between the two genders could justify the explainable discrimination present in the dataset, in fact if females work less than males it is correct for them to have a lower probability to have an income

higher than 50K, but not that high like we have seen in Figure 1.
So, we can conclude that part of the discrimination present in historical data is bad because even if males and females have similar merits there is an high bias towards males to have an income higher than 50K.
Finally, to conclude our data exploration analysis on the raw training set we decided to plot the graphs in Figure 4 and Figure 5.

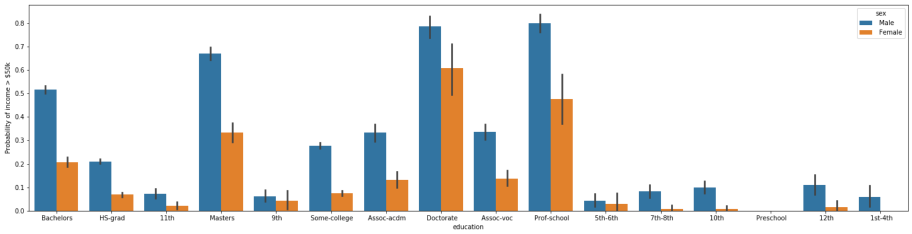

Figure 4: Probability of income higher than 50K for males and females in the raw training set partitioned by education level.

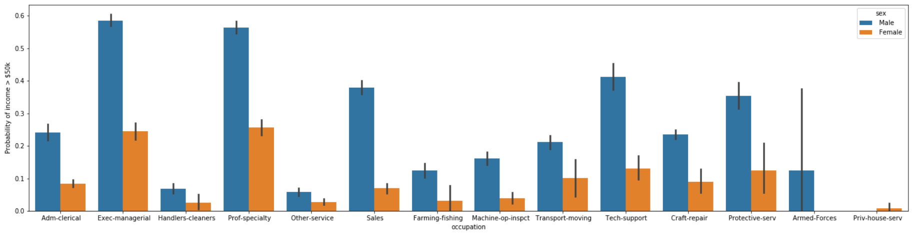

Figure 5: Probability of income higher than 50K for males and females in the raw training set partitioned by occupation type.

From Figure 4 it is possible to see that the bias towards males is present in every education level and that there are some education levels where this bias is excessive, for example in bachelors, masters and professional school education levels.

From Figure 5 it is possible to see that the bias towards males is present in every occupation type and that there are some occupation types where this bias is quite noticeable, for example executive manager, professional speciality, sales, technical support and protective service occupations.

After we performed these data exploration analysis on the training set we checked that the test set followed the same distribution. We didn't include these analysis in our report for space constraints.

3.2 Fairness metrics choice

In order to measure fairness we selected several metrics. We opted for three different metrics:

- Statistical parity: it's the difference between the probability of being positive for a person belonging to the majority group versus the probability of being positive for a person belonging to the minority group. These two numbers should be the same, this ensures that both majority and minority groups have the same probability of being classified positively;
- False positive/False negative ratios: these two metrics allow to measure the ratio between the false positive/false negative values of the majority/minority group. This two metrics allow to measure how much the model misbehaves for the two groups. When the ratios of both the majority and minority are the same we can assess that the model has basically the same performance for the two groups;
- Conditional statistical parity: it's like the statistical parity but the groups are partitioned with respect of a common characteristic (e.g: we can evaluate statistical parity on all the members of the majority/minority group that have the same instruction level). It is a more fine-grain metric that permits to evaluate fairness more precisely.

3.3 Fairness evaluation on raw dataset

Metric	Value
Statistical parity	0.20247105907754015

Table 1: Statistical parity evaluated on raw dataset.

The statistical parity value on Table 1 is compatible with what was observed on Figure 1. A positive number highlights a bias towards males. The number is well above our acceptance range which is between -0.1 and 0.1, this means that males have way higher probability of being classified above the 50K.

Conditional statistical parity	
Education level	Value
9th	0.01731201268666579
5th-6th	0.012759170653907498
11th	0.05198867745201347
7th-8th	0.07439419054384719
10th	0.09223984445421159
Preschool	0.0
1st-4th	0.0594059405940594
HS-grad	0.14176262390918404
Some-college	0.20273931008204238
12th	0.09476352524480867
Assoc-voc	0.19769676280038428
Assoc-acdm	0.2015503875968992
Bachelors	0.30833006359560744
Masters	0.3344112828549041
Doctorate	0.17738799844472286
Prof-school	0.32103359173126617

Table 2: Conditional statistical parity by education level on raw dataset.

Regarding the conditional statistical parity on education (Table 2) we can see that there's a consistent bias towards males, except in the "Preschool" level where there is no discrimination at all. This can be related to the fact that there are no data items with that education level. The values we have obtained are compatible with what was observed in Figure 4.

Conditional statistical parity	
Occupation	Value
Other-service	0.029871914413309122
Farming-fishing	0.09375
Handlers-cleaners	0.04306383810734187
Adm-clerical	0.15642949993906652
Sales	0.3100923748912524
Machine-op-inspct	0.1236087932006322
Priv-house-serv	-0.007575757575757576
Craft-repair	0.1443526066350711
Prof-specialty	0.30732097111948886
Tech-support	0.2814080327663309
Transport-moving	0.11064111037673496
Protective-serv	0.22920393559928443
Exec-managerial	0.34101592067416864

Table 3: Conditional statistical parity by occupation type on raw dataset.

The same statements are valid for the conditional statistical parity on the occupation attribute (Table 3). Interestingly there is a small reverse discrimination for the "Priv-house-serv" that can be explained by the fact that there are no males in that occupation type. The values we have obtained are compatible with what was observed in Figure 5.

3.4 Predictive model selection

After we have performed data exploration analysis and we have assessed fairness on the raw training set, we trained a predictive model on the unfair data to see if the bias was present in the predictions of the model as well.
Since the choice of the model was out of the scope of this project we just selected the model with the best accuracy on the test set.
We tried several models, such as decision trees and support vector machines and finally we found that the logistic regression was the model that performed best on this dataset, in fact we obtained an accuracy of about 84% on the test set.

3.5 Data exploration of unfair predictions

After we trained the logistic regression on the unfair dataset we got an accurate model, but accuracy doesn't imply fairness. So, we decided to perform the same analysis we performed on the ground truth labels of the unfair training set on the predictions of our model. In this way it is possible to see if the model is fair or unfair, but since it has been trained on an unfair training set without implementing fairness aware strategies, we expect the predictions to be unfair as well.

From Figure 6 it is possible to observe that the model carries the bias that was present in the training set as expected, in fact there is an high bias towards males to have an income higher than 50K even in the predictions of the model. So, we can state that if a model with no fairness aware strategies implemented is trained on an unfair dataset, it is highly probable that its predictions will be unfair too.

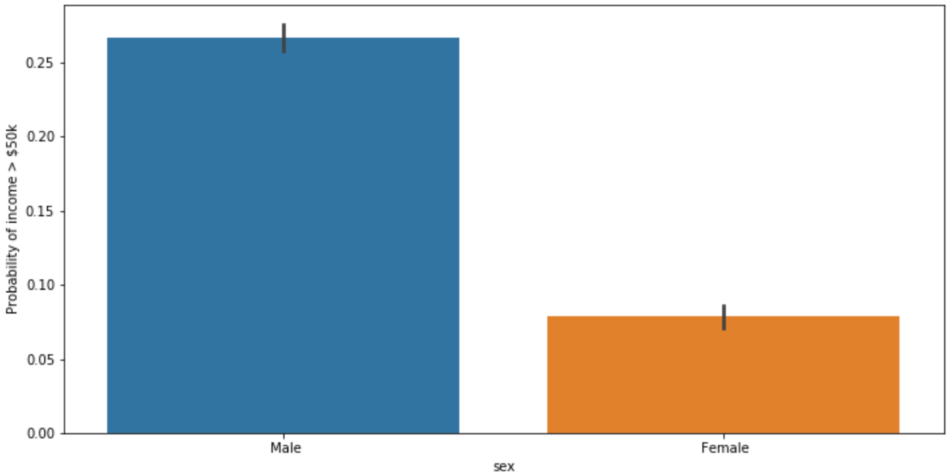

Figure 6: Probability of income higher than 50K for males and females on the predictions (test set) of the model trained on the raw training set.

In Figure 7 it is possible to observe that the bias towards males still occurs in every education level, while in Figure 8 it is possible to observe that it still occurs in every occupation type too.
It is interesting to note that the bias is higher in the education levels (professional school, bachelors and masters) and occupation types (professional speciality, executive manager, technical support, sales and protective service) where there was an high bias in the raw training set.

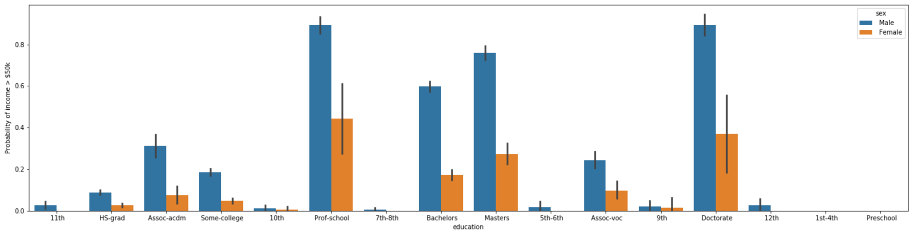

Figure 7: Probability of income higher than 50K for males and females on the predictions (test set) of the model trained on the raw training set partitioned by education level.

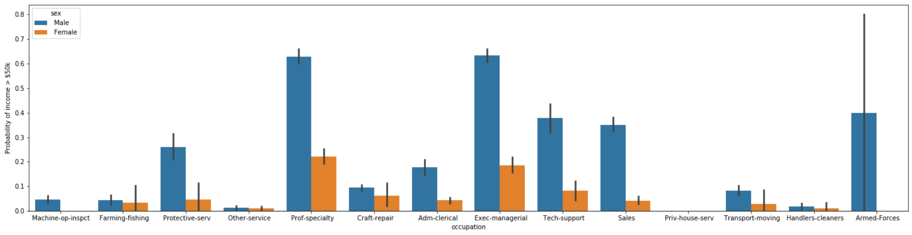

Figure 8: Probability of income higher than 50K for males and females on the predictions (test set) of the model trained on the raw training set partitioned by occupation type.

Since we found that our predictive model carries the bias that was present in the training set, we decided to try to implement some fairness aware strategies, in such a way to build a fair model. In fact, if a model is fair, it provides fair predictions even if the data which is applied to is unfair.

3.6 Fairness evaluation on unfair predictions

Metric	Value
Statistical parity	0.1880813063081353

Table 4: Statistical parity evaluated on the predictions of the model trained on the raw training set.

On Table 4 it is possible to observe that the statistical parity is compatible with what we observed in Figure 6. Basically, the statistical parity shows a bias towards males in the predictions of the model as well.

On Table 5 the false positive ratio is almost acceptable, this metric tells us that males and females have almost the same number of false positive.
The false negative ratio, instead, should be improved since the two values are very different.

Metric	Value	
	Males	Females
False positive ratio	0.2724539646749342	0.24802110817941952
False negative ratio	0.16026254615069055	0.059513074842200184

Table 5: False positive/False negative ratios of the predictions of the model trained on the raw training set.

4 Fair dataset

4.1 Our approach for achieving fairness

To try to make our model fair we opted for the preferential resampling fairness aware strategy. This is a pre-processing technique that changes the training set in such a way the it becomes fair on a specific protected attribute, the gender in our case. Before applying this technique the probability to have an income higher than 50K was about 0.31 for males and 0.11 for females. The preferential resampling simply ranks males and females per probability to have an income higher than 50K and then it deletes an amount of males that are closest to the right of the decision boundary and it duplicates the same amount of males that are closest to the left of the decision boundary, vice versa for females. This process is applied while the probabilities of the two groups converges to the mean of the two, that is 0.21 in our case.
It is important to note that we applied this technique with two modifications of the original algorithm proposed by the teacher:

1. the ranker has been retrained before each application of the algorithm. In this way every time we change the dataset we are sure to have found the best accurate decision boundary;

2. it has been used a simplified version of the algorithm that doesn't divide the dataset in partitions based on an explanatory attribute (e.g. education level or occupation type). We decided to opt for this simplified version because it took a lot of effort to just implement this version and we didn't have time to improve it.

4.2 Data exploration on fair dataset

In Figure 9 it is possible to observe that the preferential resampling fairness aware strategy has adjusted the bias that was present in the unfair dataset (Figure 1) by removing the bad discrimination that was included on it. It is possible to observe that now males and females have the same probabilities to have an income higher than 50K, so we can say the dataset is now fair on the selected protected attribute.

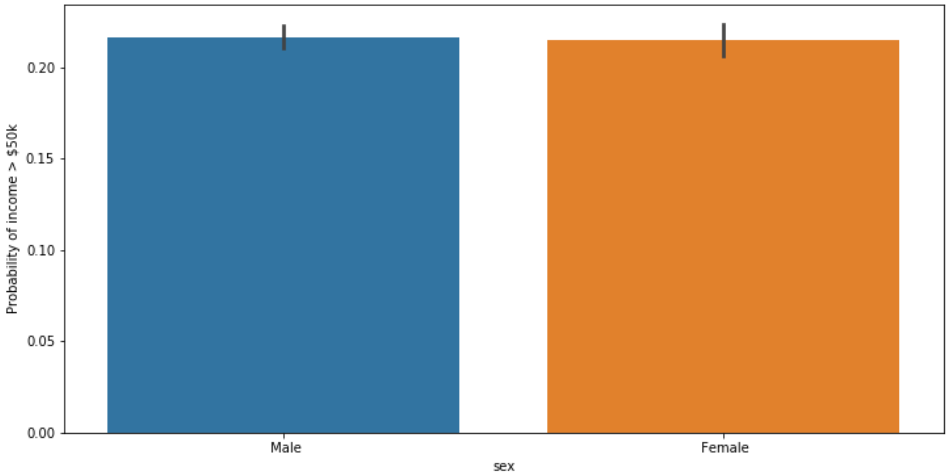

Figure 9: Probability of income higher than 50K for males and females in the modified training set.

Since preferential resampling is a pre-processing technique it is possible that its application involves some changes in the distribution of the dataset, so we decided to perform the same data analysis we performed on the unfair training set to the fair training set too, in order to see if the process has changed the distribution of the dataset.

In Figure 10 and Figure 11 it is possible to observe that males and females still have the same merits in terms of years of education and similar merits in terms of quantity of work per week, so it seems that the application of the algorithm didn't change too much the dataset on these two attributes.

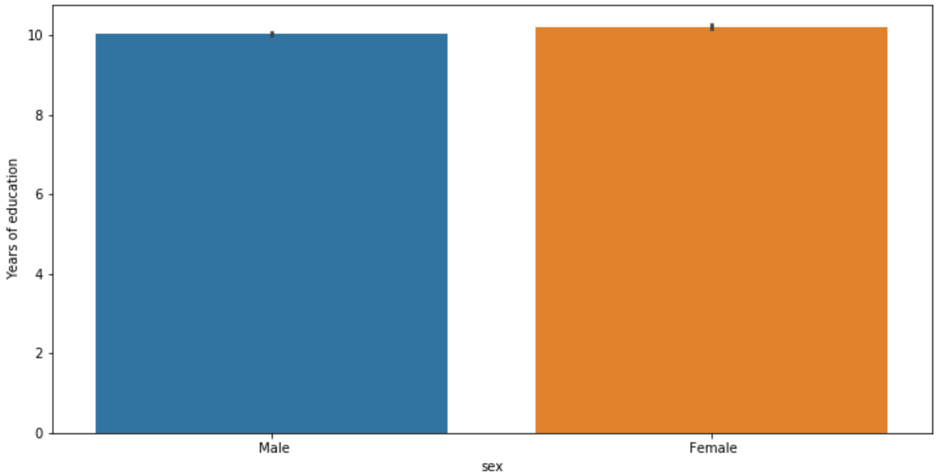

Figure 10: Average of years of education for males and females in the raw training set.

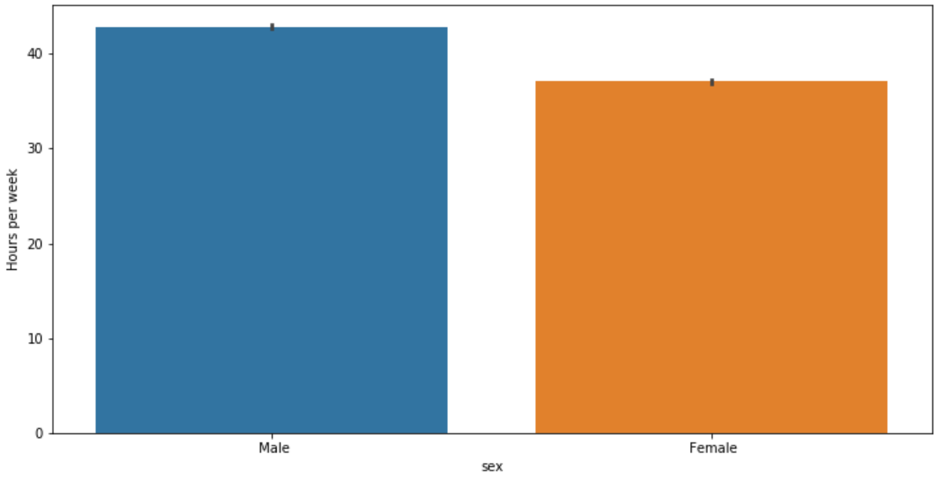

Figure 11: Average of hours of work per week for males and females in the modified training set.

In Figure 12 it is possible to observe that the technique has approximately adjusted the bias for each level of education, but in some cases it has introduced a reverse discrimination in favor of females. It is interesting to note that the reverse discrimination occurs in the education levels where there was an high bias towards males in the unfair training set. For example in masters, bachelors and professional school there is a large bias towards females to have an income higher than 50K. In Figure 4 we observed that there were an high bias towards males in these specific education levels.

In Figure 13 it is possible to observe that the technique has approximately adjusted the bias for each level of occupation, without introducing any cases of reverse discrimination, even in the occupation types (executive manager, professional speciality, sales, technical support and protective service) where there was an high bias towards males in the unfair training set.

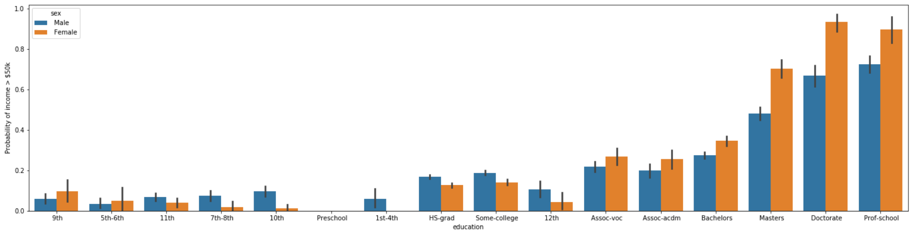

Figure 12: Probability of income higher than 50K for males and females in the modified training set partitioned by education level.

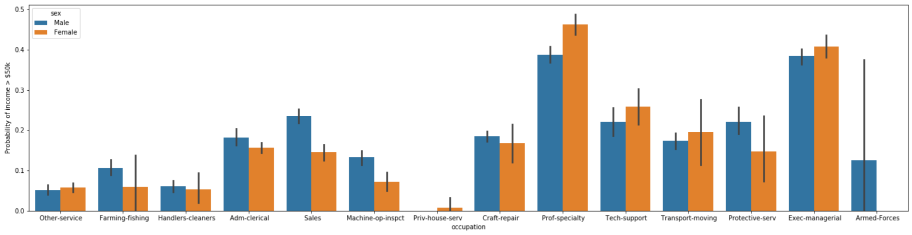

Figure 13: Probability of income higher than 50K for males and females in the modified training set partitioned by occupation type.

The technique has probably introduced reverse discrimination because we used a simplified version of the fairness aware method, which doesn't divide the dataset in partitions on the explanatory attribute.

We are sure that it is possible to obtain better results trying to obtain the same probability to have an income higher then 50K for both males and females for every level of education or occupation. If this more complicated version of the algorithm is used, the problem moves on the choice of the best explanatory attribute for the application of the technique. In fact only one explanatory attribute or a single attribute that is combination of them can be used in this method and if we make partitions on education levels it could be possible that then the bias is still present in the occupation types.

4.3 Fairness evaluation of fair dataset

Table 6 shows how the statistical parity decreased on the fair dataset. The new value is well inside our acceptance range which confirms the effectiveness of the preferential resampling.

Metric	Value	
	Unfair dataset	Fair dataset
Statistical parity	0.20247105907754015	0.0017674723522783098

Table 6: Confrontation of statistical parity on unfair and fair datasets.

Conditional statistical parity		
Occupation	Unfair dataset	Fair dataset
Other-service	0.029871914413309122	-0.0053641446296060075
Farming-fishing	0.09375	0.04801407742584213
Handlers-cleaners	0.04306383810734187	0.00810011376564277
Adm-clerical	0.15642949993906652	0.026162445348717284
Sales	0.3100923748912524	0.08928847044296959
Machine-op-inspct	0.1236087932006322	0.06024664329281834
Priv-house-serv	-0.007575757575757576	-0.008264462809917356
Craft-repair	0.1443526066350711	0.016205028221049605
Prof-specialty	0.30732097111948886	-0.07530299392116474
Tech-support	0.2814080327663309	-0.038000787091696187
Transport-moving	0.11064111037673496	-0.021374612627766548
Protective-serv	0.22920393559928443	0.07421047321158308
Exec-managerial	0.34101592067416864	-0.02474963745518599

Table 7: Confrontation of conditional statistical parity evaluated on occupation attribute on unfair and fair datasets.

Table 7 shows us that the fairness has been achieved for every occupation type since the values of the metric always stay within our acceptance range. We can also see a few cases of slightly reverse discrimination since the number is negative, Figure 13 shows that effectively.

The comparison with the old values of the metric is interesting, we can notice how all the cases of discrimination have been corrected, especially the rows in bold that are the most notable cases of discrimination have been fixed effectively by our method.

Conditional statistical parity		
Education level	Unfair dataset	Fair dataset
9th	0.01731201268666579	-0.03554778554778555
5th-6th	0.012759170653907498	-0.01634615384615385
11th	0.05198867745201347	0.027135384403606584
7th-8th	0.07439419054384719	0.05600071225071225
10th	0.09223984445421159	0.08303831327196916
Preschool	0.0	0.0
1st-4th	0.0594059405940594	0.0594059405940594
HS-grad	0.14176262390918404	0.04147966030845049
Some-college	0.20273931008204238	0.04662799314054972
12th	0.09476352524480867	0.06252467607049281
Assoc-voc	0.19769676280038428	-0.049322623980624786
Assoc-acdm	0.2015503875968992	-0.05761158640135858
Bachelors	0.30833006359560744	-0.07141174717668936
Masters	0.30833006359560744	-0.22130080985026512
Doctorate	0.17738799844472286	-0.2664735287434674
Prof-school	0.32103359173126617	-0.172545067741574

Table 8: Confrontation of conditional statistical parity evaluated on education level attribute on unfair and fair datasets.

We cannot say the same for the education levels, what we see in Table 8 is congruent with the plot in Figure 12, among the most excessive cases of discrimination only the Bachelors have been corrected. The Master education level sees basically a flipped situation where a prominent reverse discrimination has been introduced. The same thing can be said for the Doctorate and Professional school education levels, as we have already seen in Figure 12. We can see that there is no discrimination for the education level of "Preschool" but it comes from the fact that the are no data items having that level of education in the training set.

4.4 Data exploration of fair predictions

After we modified the dataset in such a way to make it fair respect to the gender attribute we trained the same logistic regression model on the fair dataset to see if the model became fair. We obtained an accuracy of 81% on the test set with this model, so the accuracy has been decreased by 0.03% due to the fact that we have changed the dataset. It is usually the case when fairness aware strategies are implemented on the dataset or in the model itself.

After we trained the model we performed the same data analysis we have performed for the predictions of the model trained on the unfair dataset on the predictions of the model trained on the fair dataset.

In Figure 14 it is possible to observe the effectiveness of the preferential resampling method, in fact the bias that was present in Figure 6 has been corrected and now the predictions are more or less fair. The gap between the two probabilities is just about 0.03 while before it was around 0.19. It is interesting to note that now there is a small bias towards females in the predictions of our model. This could be due to the fact that we have applied a simplified version of the preferential resampling that has introduced some cases of reverse discrimination in favor of females in some education levels.

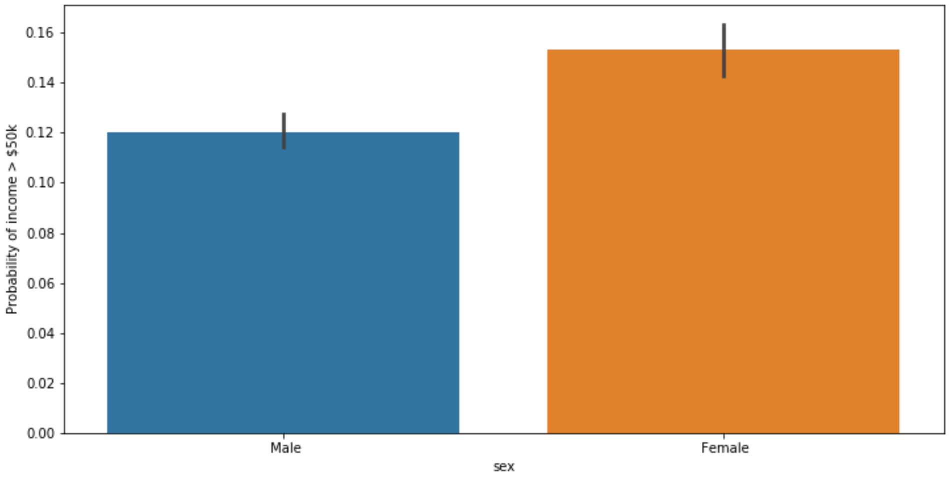

Figure 14: Probability of income higher than 50K for males and females on the predictions (test set) of the model trained on the modified training set.

In Figure 15 it is possible to observe that the bias has been corrected for each level of education. There is still some bias towards males on professional school, masters and doctorate education levels like it was in Figure 7 but in this case we can say that is due to the fact that in the test set there is a large standard deviation related to females in these educational levels. This basically means that we have only few females with these levels of education on the test set and that's why we see this large bias on the predictions of the model.

In Figure 16 it is possible to observe that the bias has been corrected for each level of occupation as well. It is interesting to note that in this case the preferential resampling has been more effective compared to the previous case, in fact there was a large bias for each level of occupation in Figure 8 that now has been corrected.

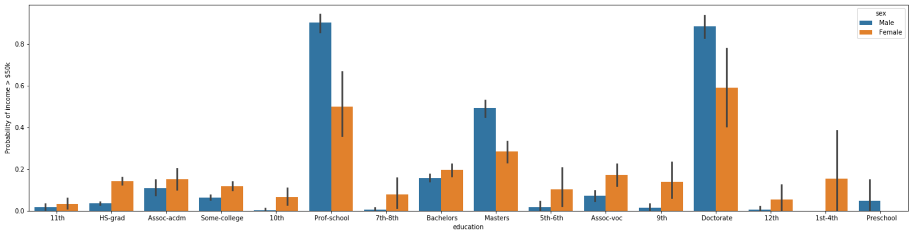

Figure 15: Probability of income higher than 50K for males and females on the predictions (test set) of the model trained on the modified training set partitioned by education level.

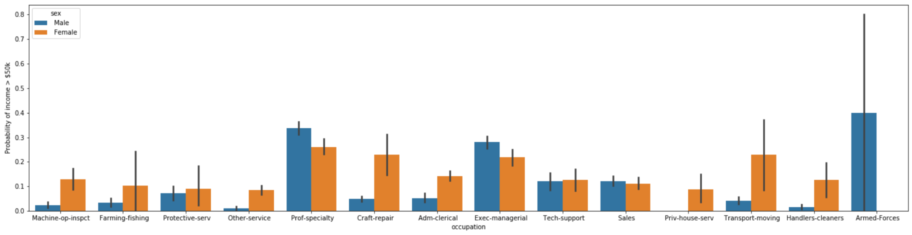

Figure 16: Probability of income higher than 50K for males and females on the predictions (test set) of the model trained on the modified training set partitioned by occupation type.

4.5 Fairness evaluation on fair predictions

Metric	Value	
	Not fair model	Fair model
Statistical parity	0.1880813063081353	-0.03265026972516913

Table 9: Confrontation of statistical parity on unfair and fair predictions.

Metric	Value			
	Unfair model		Fair model	
	Male	Female	Male	Female
False positive ratio	0.27245	0.24802	0.14571	0.48440
False negative ratio	0.16026	0.05951	0.23732	0.04144

Table 10: Confrontation of False positive/False negative ratios on unfair and fair predictions.

Regarding the False positive ratios in Table 10, we can see how the value for the males decreased while the value for the females greatly increased as expected. In

fact, since the model now is fair it will classify positively way more females than what expected by the test set, hence the high false positive ratio.
The same process goes for the rise of the false negative ratio for the males, our model penalizes man in order to be fair.

Looking at the statistical parity in Table 9 we can notice that now the predictions are more balanced among the two genders as confirmed by the plot in Figure 14. Furthermore, it is possible to observe that a small reverse discrimination is present in the predictions of our model, like we have already seen in the previous section.

Finally, it is interesting to note that usually a fairness project should be focused on improving only one metric because the improvement of a metric is often related with the aggravation of a complementary metric. In our case we obtained good results on the statistical parity but large gaps in the false positive and false negative ratios, as expected.

5 Conclusions

We are satisfied with the results we achieved. The method we applied successfully removed discrimination from the dataset and allowed us to train a fair model. Table 8, Table 7 and Table 10 show how the discrimination changed for each education/occupation type. Despite a few cases of reverse discrimination we can notice how the fairness has been fixed at almost every level of education and type of occupation.

Even though we applied a simplified version of the preferential resampling we still managed to achieve very good results without loosing too much accuracy, so we are confident that introducing the partition by the explanatory attribute will result in better outcomes and a better accuracy.

Finally, since we didn't have enough time to perform all the experiments we wanted to perform, we have a few margins of possible improvements that could be future works:

- implementing the preferential sampling dividing in partitions by an explanatory attribute;
- implementing twin tests;
- implementing a more fine-grain fairness adjustment by taking into consideration also the race of the subjects;
- implementing a logistic regression model with learning constraints, in order to make the model fair without modifying the data.