Additional insights

Katarzyna Gunia, Krystian Walewski

November 2023

1 Distribution of features

In this section, one can find an overview of features in the Adult dataset.

Plots 1 and 2 show distribution and boxplot of age, respectively. We can see that the range age seems rational and that the majority of record in the dataset represent people between age 25 and 47.

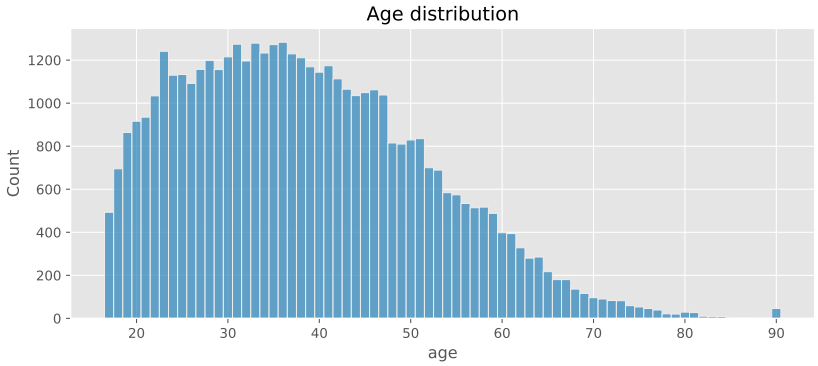

Figure 1: Distribution of age.

Figures 3 and 4 describe the distribution of number of hours worked per week. Also, this time the value range seems to be possible and that the majority of people work 40 hours per week.

The next plot 5 justifies why we decided to introduce a new feature "from_USA" that merges all rarely occurring native countries. Figure 6 shows the distribution of that new feature, which is still unbalanced. We also considered merging countries from the same continent, but in that case there would still be very rarely occurring categories.

Plot 7 reflects the number of records representing each occupation. Figure 8 shows the distribution of workclass.

The next plot 9 represents the number of people with each education level in the dataset. The most popular categories are: high-school graduate, collage graduate and bachelor.

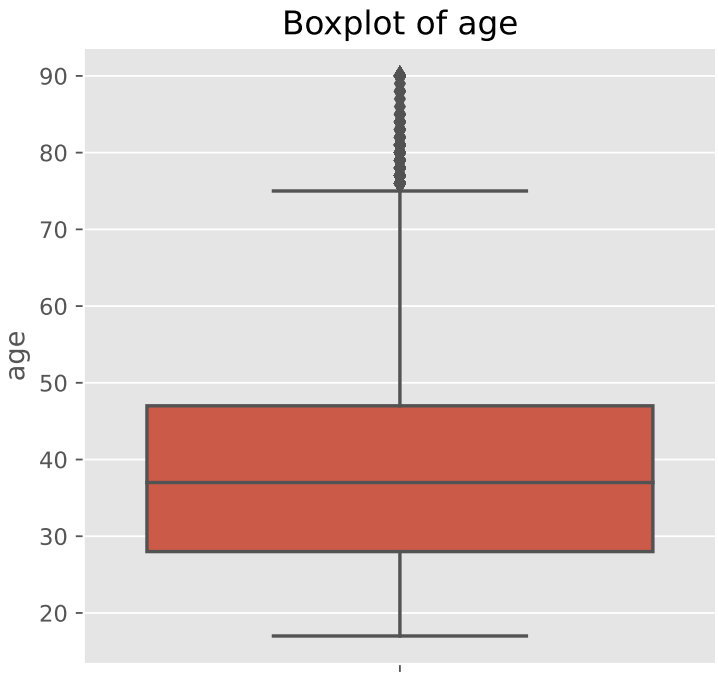

Figure 2: Boxplot of feature age.

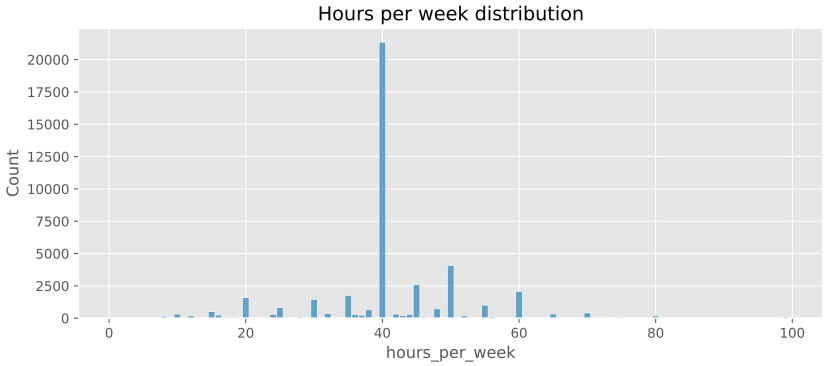

Figure 3: Distribution of number of worked hours per week.

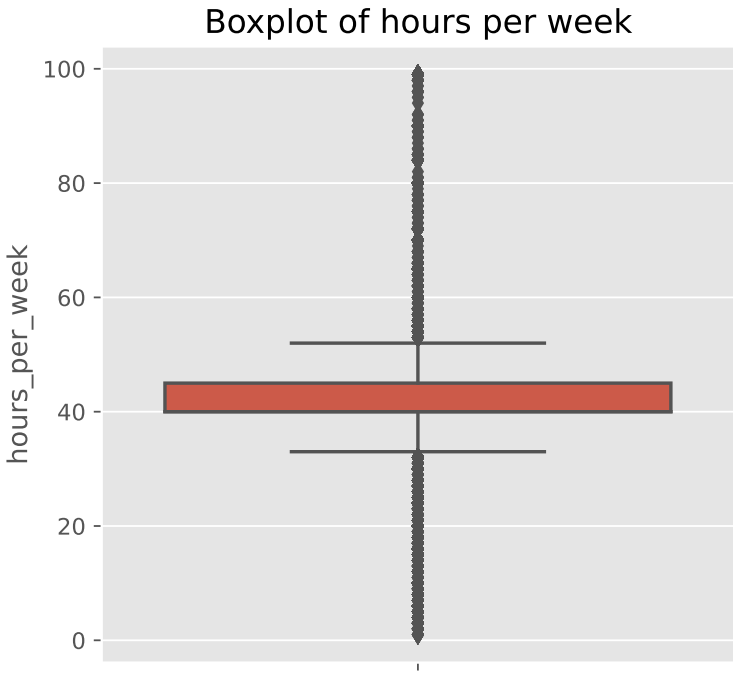

Figure 4: Boxplot of number of worked hours per week.

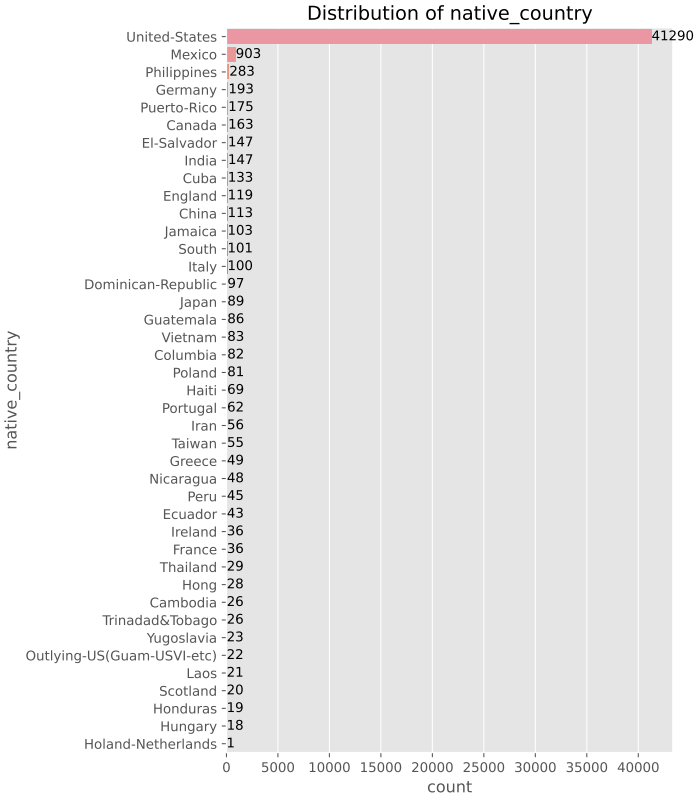

Figure 5: Distribution of native country.

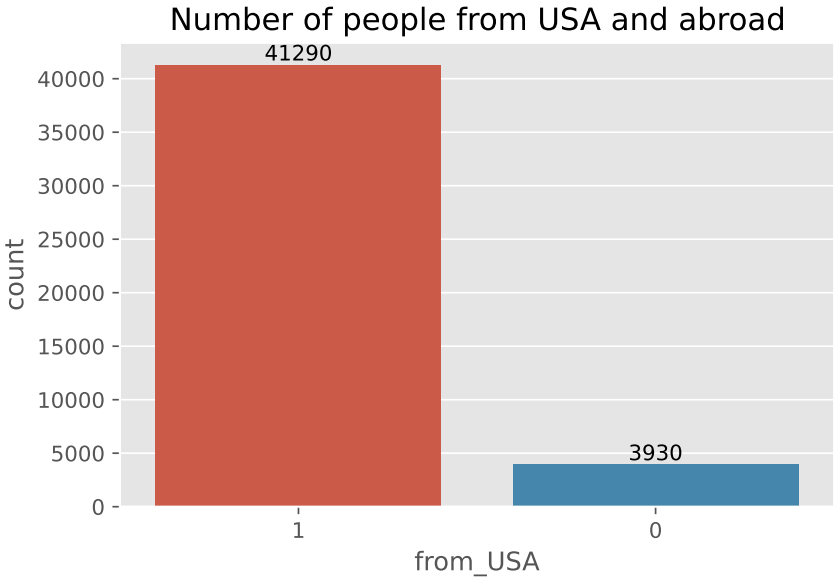

Figure 6: Distribution of feature from_USA.

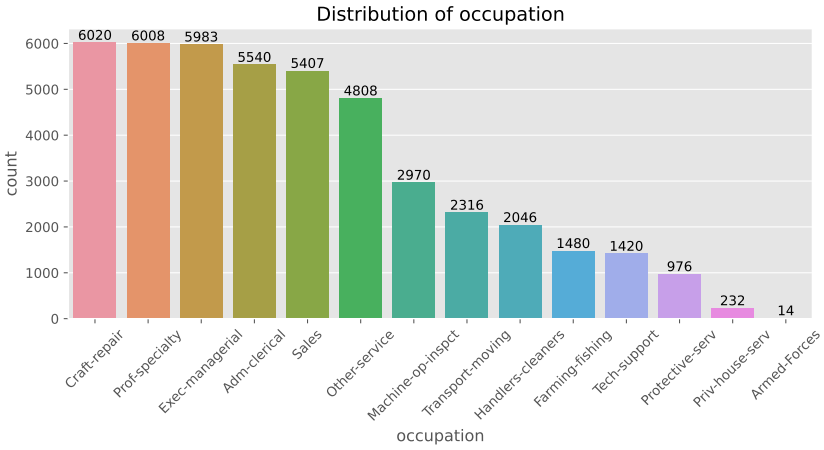

Figure 7: Distribution of occupation.

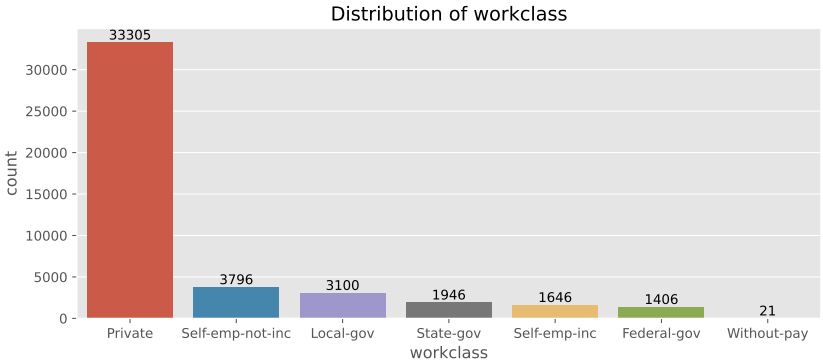

Figure 8: Distribution of work class.

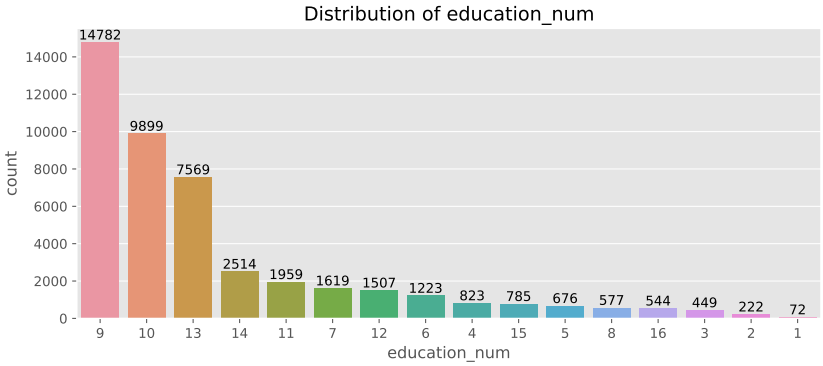

Figure 9: Distribution of education-num.

Figures 10 and 11 describe features marital-status and relationship. Approx. 12% of persons are married and 8% were never married. Those numbers match the trend in the relationship variable. Where category husband and not-in-familly are the two most popular.

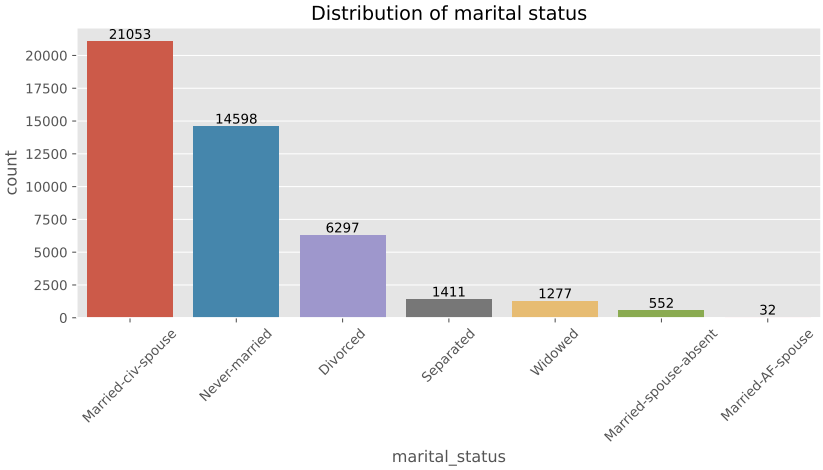

Figure 10: Distribution of marital status.

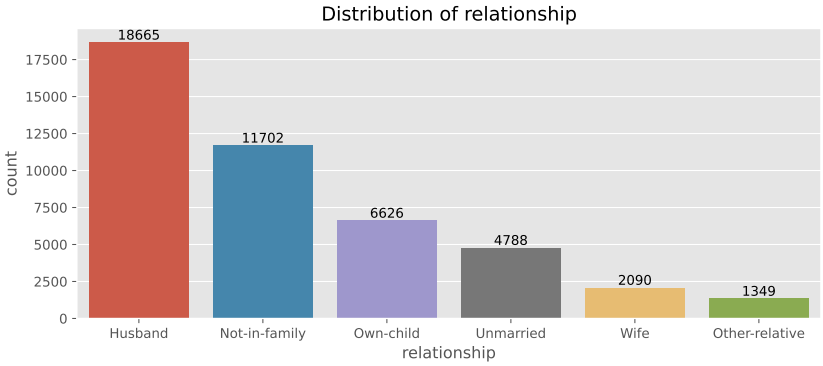

Figure 11: Distribution of relationships.

Distribution of race displayed on the plot 12 show that we have definitely much more white people compared to other races in the data. We also have definitely more men than women, as shown on the plot 13.

Figure 14 reflects the distribution of the mysterious feature, fnlwgt. Plots 15 and 16 shown the distributions of capital gain and loss, respectively. One can notice that the last two features are mostly zeros.

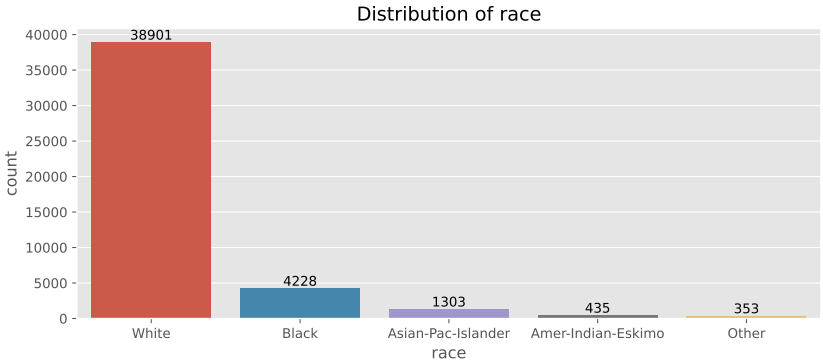

Figure 12: Distribution of race.

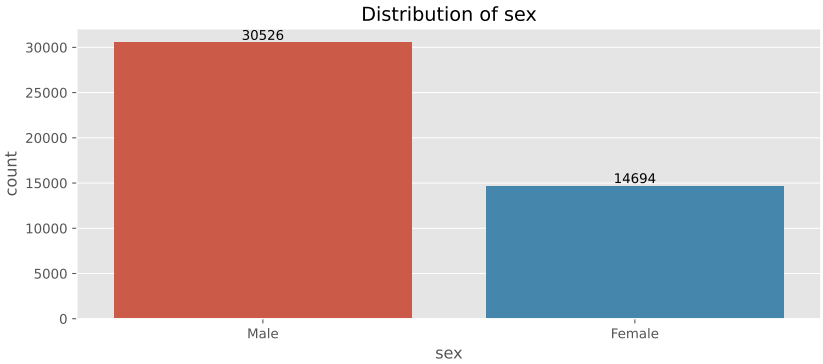

Figure 13: Distribution of sex.

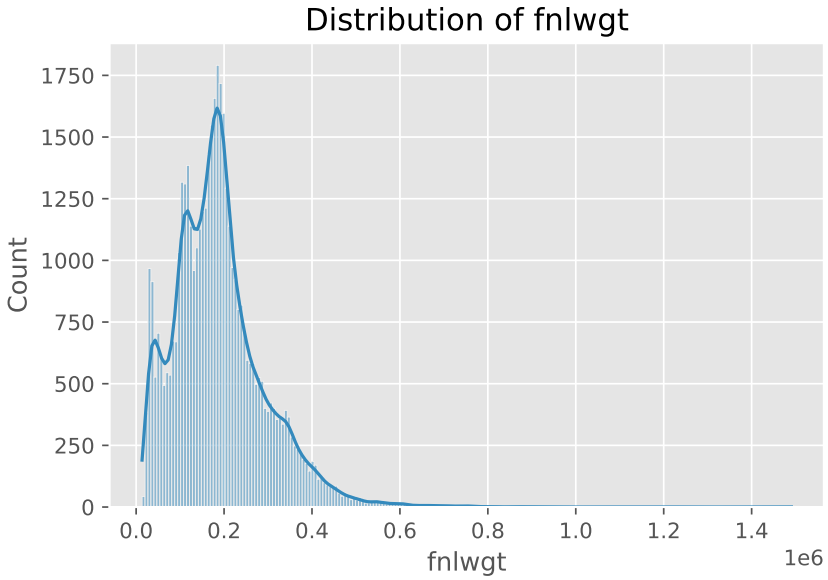

Figure 14: Distribution of feature fnlwgt.

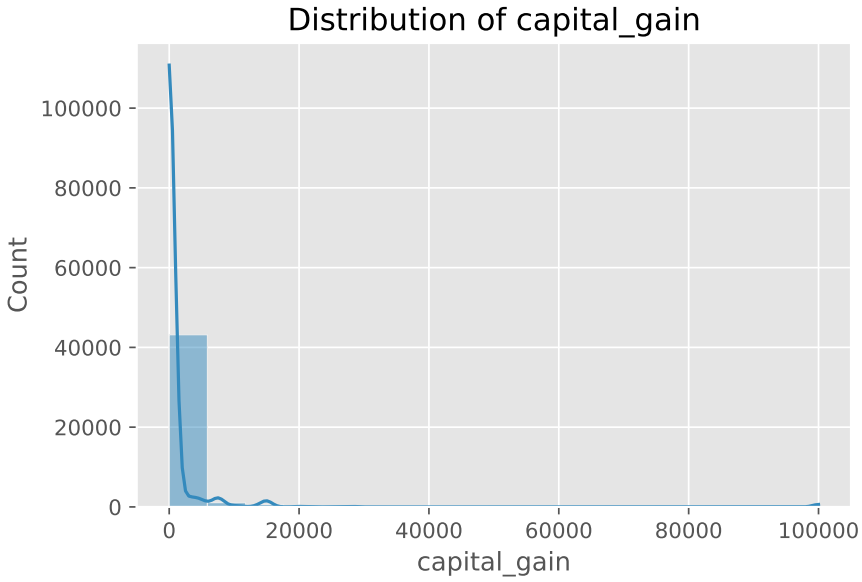

Figure 15: Distribution of capital gain.

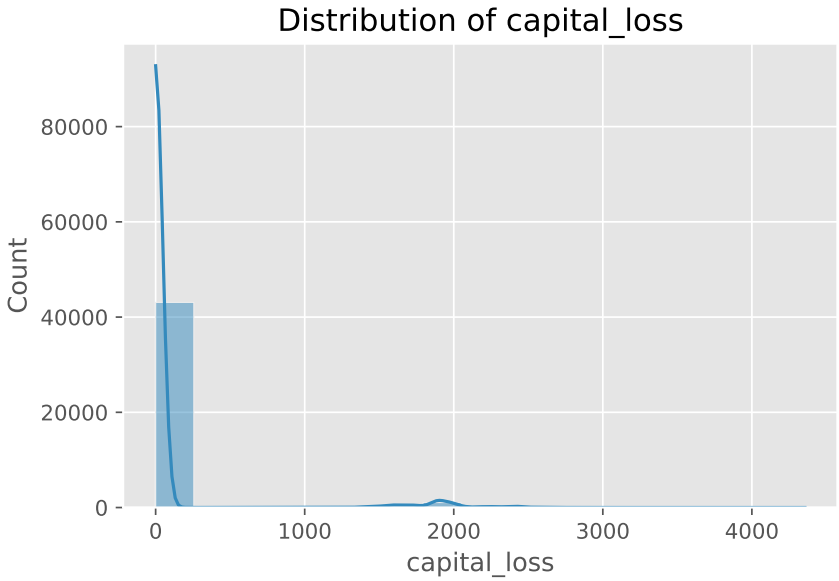

Figure 16: Distribution of capital loss.

2 Correlation

In table 1 one can find a correlation between income and remaining features. In addition to Person's correlation, we also performed Chi-squared test, to check if categorical variables are correlated with each other. We got the following results. All features are correlated with each other, apart from race and from_USA.

3 Distributions within groups

Here are some features' distributions with respect to the target variable — income, which were not shown in the main report:

- plot 17 — workclass,
- plot 18 — sex,
- plot 19 — race,
- plot 20 — fnlwgt,
- plot 21 — from_USA.

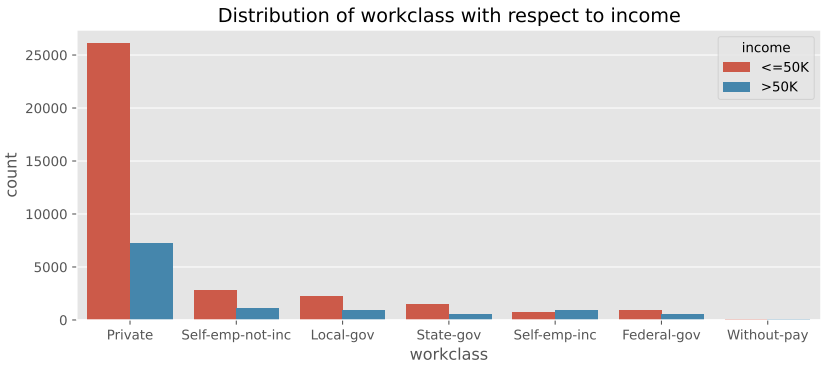

Figure 17: Distribution of workclass by income.

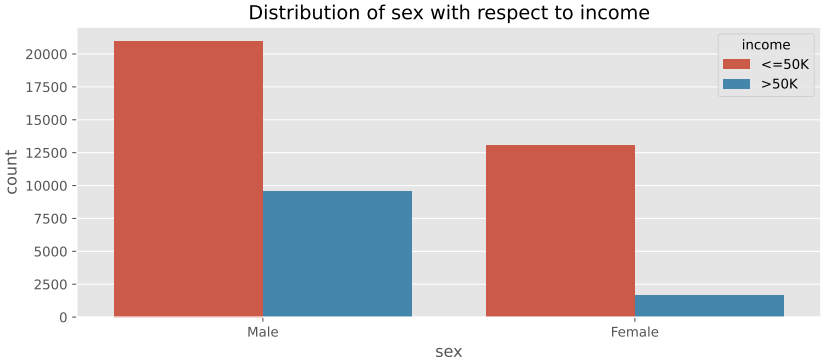

Figure 18: Distribution of sex by income.

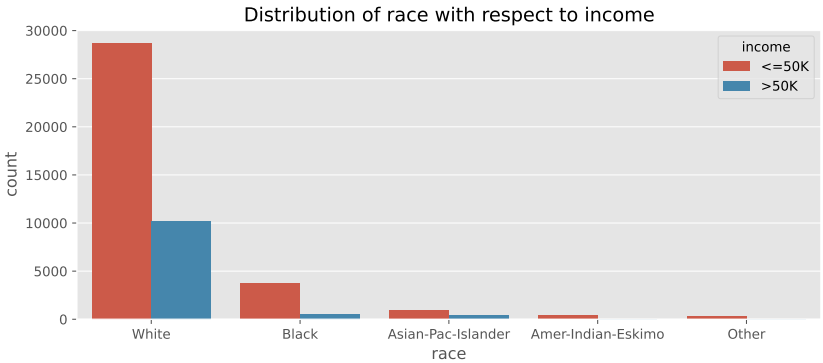

Figure 19: Distribution of race by income.

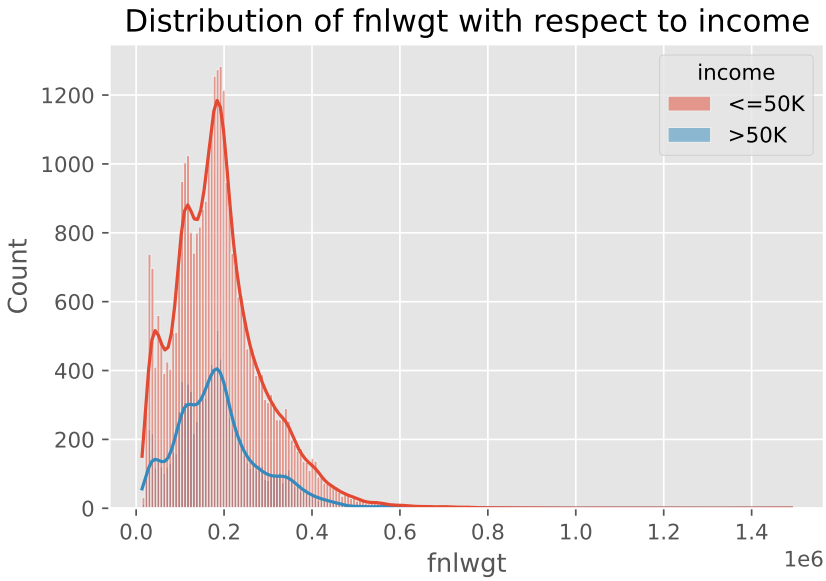

Figure 20: Distribution of fnlwgt by income.

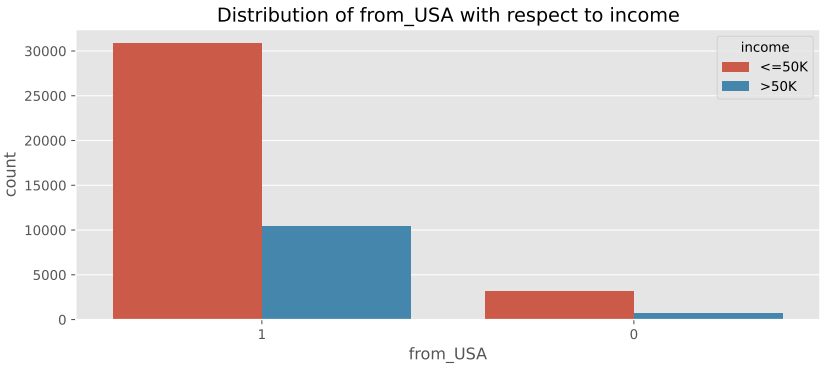

Figure 21: Distribution of from_USA by income.

	income
age	0.237074
workclass	0.015665
fnlwgt	-0.007325
education_num	0.332769
marital_status	-0.192708
occupation	0.049835
relationship	-0.253526
race	0.070837
sex	0.215730
capital_gain	0.221046
capital_loss	0.148698
hours_per_week	0.227210
from_USA	0.038907

Table 1: Pearson correlation of income and remaining features.

4 Hyperparameter tuning

Plot 22 presents the accuracy as a function of number of neighbors in KNN model. We looked for the optimal number of neighbors between 1 and 45 using grid search and stratified cross validation. It turned out that 38 is the optimal number, however we can see on the plot that the curve starts to flat around $n = 15$.

Below is the list of optimal models' parameters chosen in the same manner, using grid search and maximizing accuracy.

- KNN
  - p = 2
  - weights = uniform
- LDA
  - solver = svd
- Random Forest
  - criterion = gini
  - n_estimators = 80
- XG Boost
  - eta = 0.3
  - max_depth = 4
  - scale_pos_weight = 1
  - n_estimators = 140

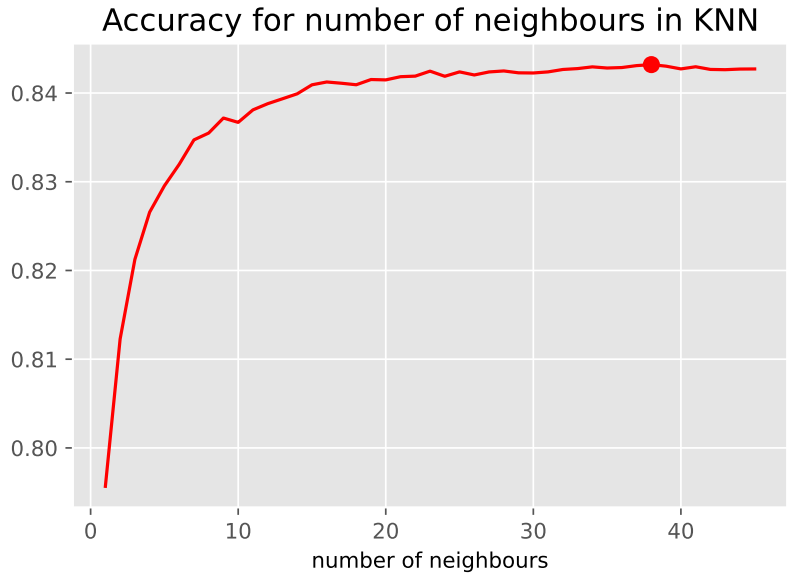

Figure 22: Accuracy of KNN model as a function of number of neighbours. Dots marks the largest accuracy for 38 neighbours.

5 Additional results

In this section, we can find the comparison of boxplots of precision and recall for different models. Plot 23 and 24 contains results for the initial models. The next two plots: 25 and 26 show precision and recall respectively for models trained on the dataset without features fnlwgt, workclass and from_USA. And the last two plots: 27 and 28 contains analogical results but for models without feature fnlwgt.

Quadratic Discriminant Analysis acts very interestingly, initially it has the best recall and the worst precision, after feature selection it has better precision than before, now it's on the similar level as other models precision, but its recall drastically drops.

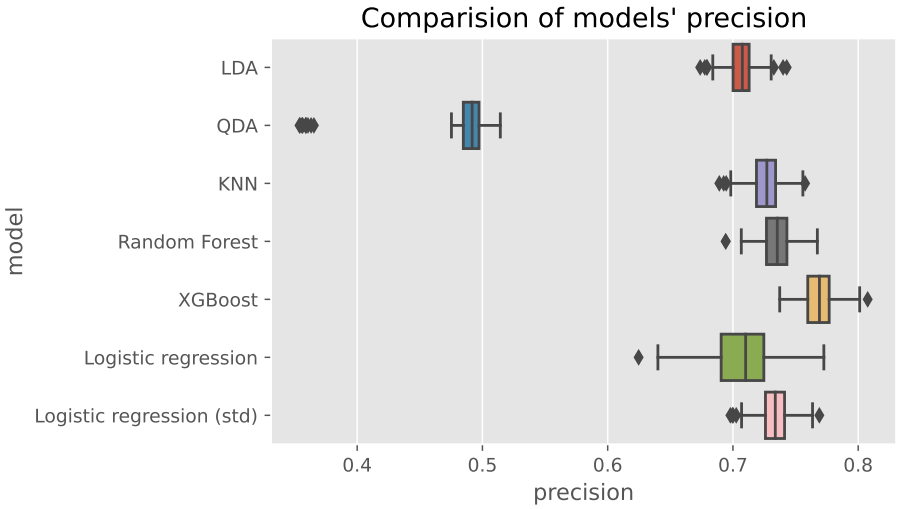

Figure 23: Boxplots of precision for considered models.

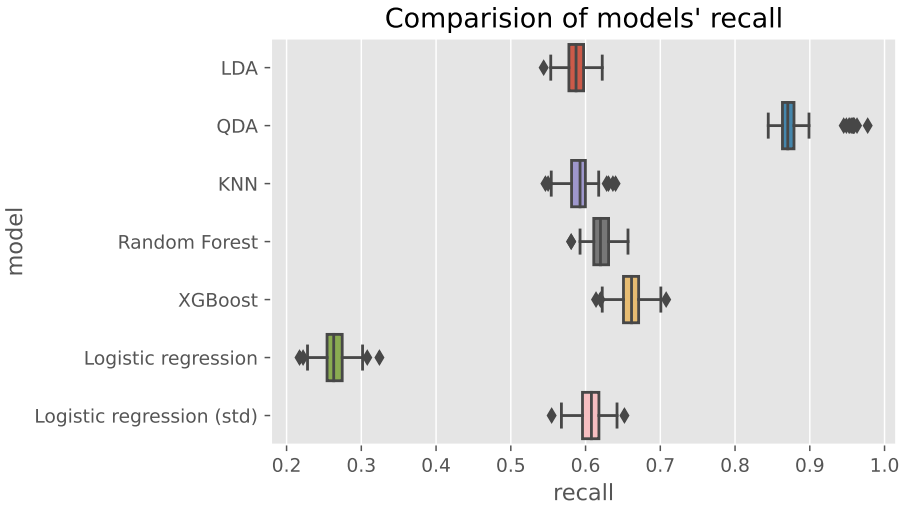

Figure 24: Boxplots of recall for considered models.

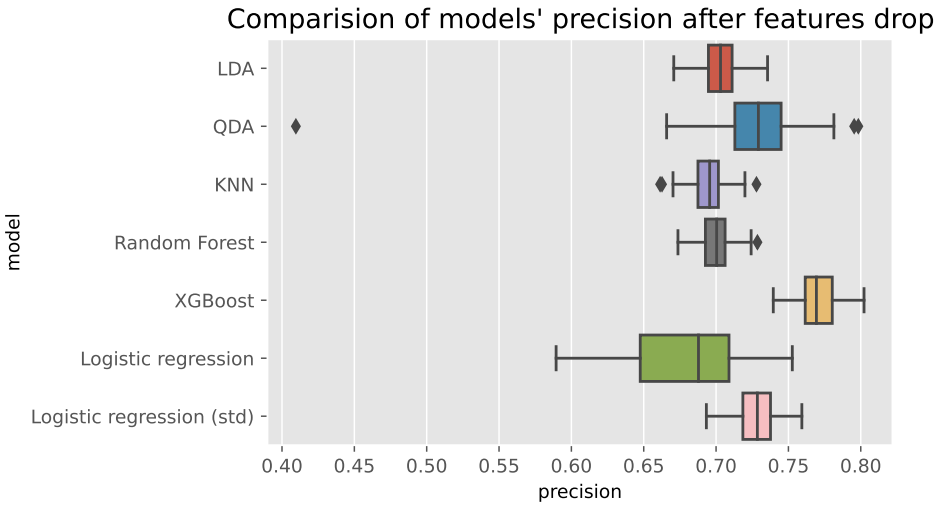

Figure 25: Boxplots of precision for considered models after dropping features: fnlwgt, workclass and from_USA.

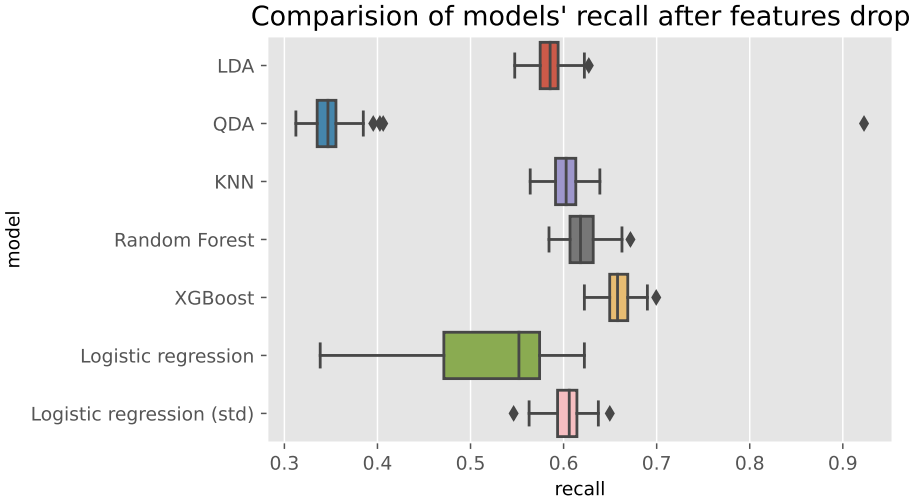

Figure 26: Boxplots of recall for considered models after dropping features: fnlwgt, workclass and from_USA.

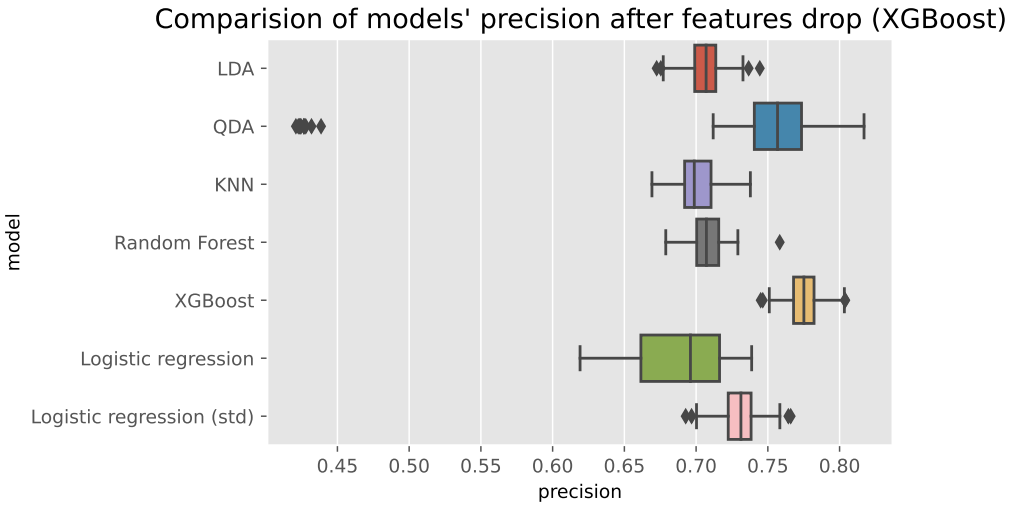

Figure 27: Boxplots of precision for considered models after dropping features: fnlwgt.

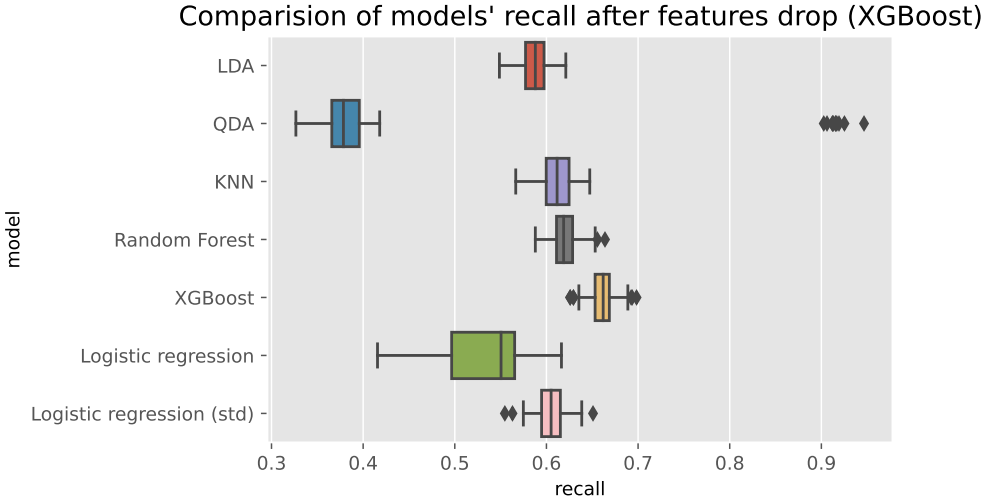

Figure 28: Boxplots of recall for considered models after dropping features: fnlwgt.