Logistic Regression Analysis of the Income Level in the US Based on Socio-Demographic Variables

Ghali Chraibi

July 6, 2022

1 Introduction and Background

In the labour market, several regulations define how wages should be set. The salary is negotiated between an employee and an employer in the private sector, whereas governmental laws regulate the salary of public employees in a more systemic way. In particular, in the United States, the Title VII of the Civil Rights Act of 1964 states that there shall be no discrimination between employees based on demographic characteristics (e.g. race, sex, national origin).

The aim of this study is to predict whether a person earn more than 50'000$ per year based on socio-demographic data.

The **null hypothesis** is that none of the socio-demographic variables below have a statistically significant impact in predicting whether a person makes over 50'000$ a year as required by US law. Hence, the **alternative hypothesis** is that at least one socio-demographic variable have a statistically significant impact in predicting whether a person makes over 50'000$ a year.

The data used for this analysis were extracted from the 1994 Census Bureau Database. It is called the Adult dataset and was gathered by Kohavi [2]. It contains demographic and economic information about US population.

In particular, we will use the following variables: *age* (Integer), *sex* (Categorical), *race* (Categorical) describing the most common races in the US with a value *Other* gathering the least frequent, *native-country* (Categorical), *education-num* (Ordinal) encoding the level of education, *marital-status* (Categorical) corresponding to the civil status and is complementary to the *relationship* (Categorical), *workclass* (Categorical) indicating whether the person works for the public or private sector or is self-employed, *occupation* describing the kind of job of the person, *capital-gain* (Integer) which is the money earned through goods and investments, *capital-loss* (Integer) which is conversely the money lost through goods and investments and finally *hours-per-week* (Integer) describing the number of hours worked per week in average.

We did not consider the following variables used by Kohavi in his analysis [2]: *education* as it is redundant with *education-num* and *fnlwgt* as this variable is an aggregation of several demographic variables and we do not know how it was computed, hence it would not be very interpretable.

2 Exploratory Data Analysis

Note that the exploratory data analysis is only performed on the training set which corresponds to two third of the total data collected.

First and foremost we performed some pre-processing on the data. We converted the response variable *upper_50K* as a categorical variable taking either value 0 (False) or 1 (True). We also removed 2399 rows (7.95% of the data) containing unknown values which leaves us with 30162 rows.

Exploring now the data, we observe that there is an imbalance in the distribution of the target variable towards people earning less than 50K per year as this class corresponds to 75.1% of the data, whereas the remaining 24.9% earn more than 50K. If this seems realistic, this may bias our model and encourage it to perform better on the majority class.

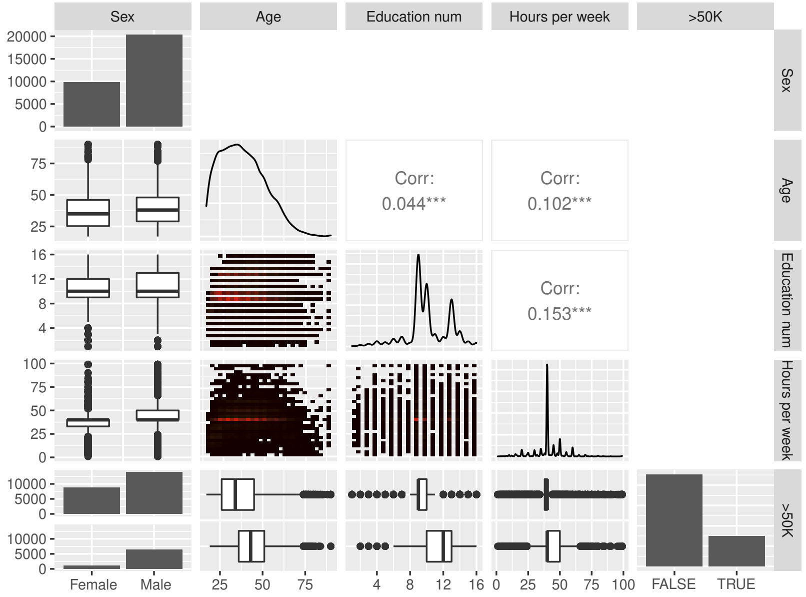

Figure 1: *Scatterplot matrix of selected variables.*

Looking at the univariate graphs on the diagonal of figure 1, we observe that we have an imbalance towards men, the latter representing approximately two third of the data. The ages are distributed more or less uniformly across the active population with slightly less people close to retirement. In regard to the level of education, three peaks can be distinguished at level 9, 10 and 13 which correspond respectively to a high school diploma, some years of college and a Bachelor's degree. Finally, we observe a large peak at 40 hours of work per week, which is not surprising as this was the recommended standard in the US.

Focusing on the bivariate graphs, we observe from the left boxplots that there are more men working slightly more hours per week, otherwise the distribution of education level and age are similar for both gender.

The tiles plot regarding the education number, the age and the hours per week does not bring more information than what we saw in the univariate analysis. The red tiles corresponding to a big concentration of data points are consistent with the peak in the univariate graphs, although these bivariate graphs still provide information that the dataset is dense. Numerically, there are no strong pairwise correlations between the three variables. However, an increase in age or education level is correlated with slightly more hours worked per week. The correlations are highly significant even though the effects are small, because a large number of datapoints are used.

Finally, we observe from the graphs in the last row that people earning more than 50K (the lower graphics) works slightly more per week, have a greater level of education, are older and are mostly men. If this sounds fair for most of these observations, the different distribution of income level in respect to the sex variable already indicates that there is a problem.

Table 1: *Summary of capital gain and loss*

	Min	1st Quartile	Median	Mean	3rd Quartile	Max
Capital gain	0	0	0	1092	0	99999
Capital loss	0	0	0	88.37	0	4356

From table 1, we see that the variables capital gain and capital loss are very sparse. The capital loss variable has a low mean and quite low value in general. On the contrary capital gain has a 0.90 quantile of 0 and a 0.95 quantile of 5013, this shows that most people did not have a capital gain, but those who benefited from it earned a lot of money.

A final note on this exploratory analysis is that the variable race could have been interesting to investigate, however there is a strong imbalance towards the *White* value.

3 Model fitting

3.1 Model definition

As we are trying to predict a binary response variable (the income level), we use the appropriate Generalized Linear Model (GLM) for binary classification, the logistic regression. A GLM is a model of the form:

$$g(Y) = \beta_0 + \sum_i \beta_i x_i \tag{1}$$

when the distribution of Y belongs to the exponential family.

In the case of the logistic regression, the link function $g(.)$ is the logit function $logit(x) = \log\left(\frac{x}{1-x}\right)$ and we do not model the the response directly, but we model instead the probability $\pi(x)$ of obtaining the value 1:

$$\pi(x) = P(Y=1|x) = \frac{exp(\beta_0 + \sum_i \beta_i x_i)}{1 + exp(\beta_0 + \sum_i \beta_i x_i)} \tag{2}$$

As logistic regression predicts probabilities, we can use the maximum likelihood estimation (MLE) to fit the model. Hence, we want to find the coefficients β that maximise the likelihood, which is equivalent to find the β that maximise the log-likelihood (as the logarithm is a strictly convex function):

$$l(\beta) = log(L(\beta)) = \sum_{i=1}^{n} (y_i log(\pi(x_i)) + (1-y_i) log(1-\pi(x_i))) \tag{3}$$

3.2 Model selection

For the model selection, we start from the model used by Kohavi [2] and perform a backward elimination. For that, we do not consider the p-values associated with each variable as it is not a good metric given our dataset. Indeed, with a sample size of over 30000 data points, any effect, no matter how small, will be significant. Therefore, the variables were removed based on their effect size.

We use the *glm* method with the binomial distribution (hence using logit link function). Note that the categorical variables are considered as sets of factors, with each factor having its own coefficient. The initial model formula is:

$$\begin{aligned} upper\_50K \sim age + race + sex + education\_num + education + native\_country + \\ marital\_status + relationship + occupation + workclass + fnlwgt + \\ capital\_gain + capital\_loss + hours\_per\_week \end{aligned} \tag{Kohavi's model}$$

We first remove the variables *education* and *fnlwgt* for the reasons explained in the first section. Then we iteratively remove the variable with the smallest effect size. Finally, we also remove the *marital_status* variable even though one of its factor has a large effect size (the value *Wife*), because the latter variable is, by design, closely correlated with the *relationship* variable. The Final model formula is as follow:

$$\begin{aligned} upper\_50K \sim age + sex + education\_num + relationship + occupation + \\ capital\_gain + capital\_loss + hours\_per\_week \end{aligned} \tag{Final model}$$

One way to compare the two models at the training level is to compute their AIC score. The Akaike Information Criterion is a method that measure how well a model fit the data it was generated from. It depends on the number of independent variables present in the model and the log-likelihood estimate. It is mathematically defined as $2K - 2ln(L)$, where K is the number of predictors in the model and L is the maximum likelihood of the model. The lower the AIC score, the better the model is able to predict the observed data with as few parameters as possible [4].

Based on this score, Kohavi's model (Kohavi's model) is slightly better (AIC of 19678) than the Final model proposed (Final model) (AIC of 19924).

4 Model assessment

For the model assessment, we use the test set pre-processed similarly to the train set.

The validity of the selected logistic model relies on some important assumptions. If the latter do not hold, it would be an indicator that the choice of model is not appropriate. The assumptions when using logistic regression are the following:

1. Binary response variable
2. Independence of the observations

3. Linear relation between the logit and the linear predictor
4. No multicolinearity between the independent variable
5. There should be no outliers

The first assumption trivially holds. The fact of having a binary response variable has in fact led to the choice of this family of GLM.

Verifying the independence of the observations is already a bit more tricky to check. However, because the data come from the Census Bureau Database, one can assume that this sample is representative of the population in the US at that time. On the other hand, it also means that this model should only work well with observations from the same population (taking a sample from another country or time period may not work well).

For the relation between the logit and the linear predictor, the plot of each dependent variable against the logit is diplayed in Figure 2. We observe in most of these plots that the data have a linear shape, but each time with a small group of outliers that has a strong impact on the odds. This could be explained by the fact that there are not enough data for the class earning more than 50K.

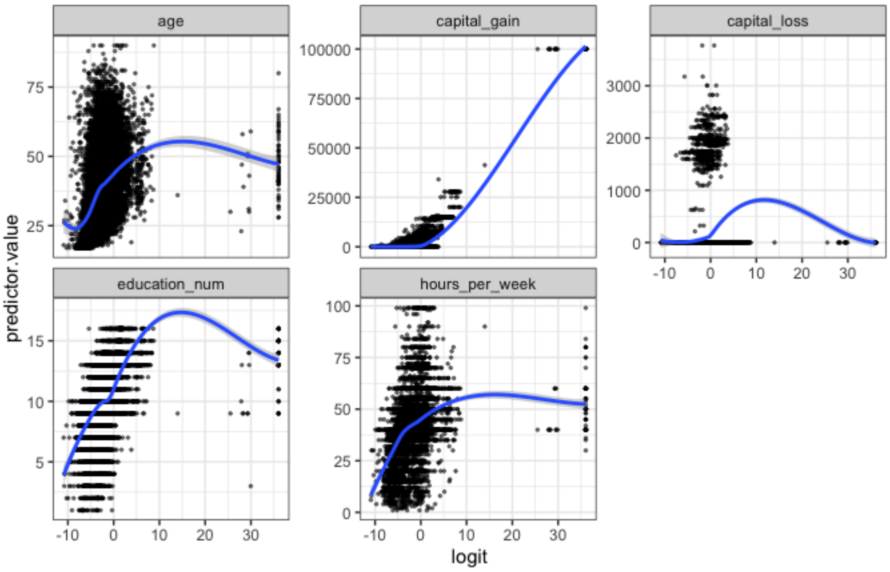

Figure 2: *Scatter plots of logit against (non-categorical) predictors*

Multicollinearity occurs when the predictors variable of a model are correlated. The Variance Inflation Factor (VIF) aims at estimating how much the variance of a coefficient is inflated due to multicollinearity within the model. In contrast with pairwise correlation, the VIF estimates the correlation of each variable with all the other variables of the model at the same time. A VIF of 5 indicates the presence of moderate multicollinearity and higher values can affect the model [1]. The Generalized Variance Inflation Factor (GVIF) is computed instead when the model contains categorical variables. To make the GVIF comparable accross the variables, the standardized GVIF of a variable is computed by taking $GVIF^{1/2Df}$ where DF is the degree of freedom of the variable. The latter hence accounts for the number of factors in Categorical variables.

Table 2: *GVIF of the Final model*

	GVIF	Df	$\text{GVIF}^{1/(2Df)}$
age	1.08	1	1.04
sex	2.78	1	1.66
education_num	1.47	1	1.21
relationship	3.05	5	1.11
occupation	1.74	13	1.02
capital_gain	1.02	1	1.01
capital_loss	1.01	1	1.00
hours_per_week	1.10	1	1.05

From the Table 2, the assumption of no multicollinearity holds for the Final model.

To verify the last assumptions that there are no outliers, we compute the Cook's distance. It allows to find the outliers that negatively influences the analysis in a set of predictors and is computed based on the residuals of the observations. More formally, the Cook's distance of an observation i is the sum of all the changes in the regression model that occurs when i is removed from it. A rule of thumb is that an observation with a Cook's distance greater than 4/N (with N being the total number of datapoints) is considered as an influential observation [3]. Observing many such outliers is an indicator that there is a problem with the logistic model.

According to Figure 3, there are some outliers in the model. It could be interesting to take them individually to understand them.

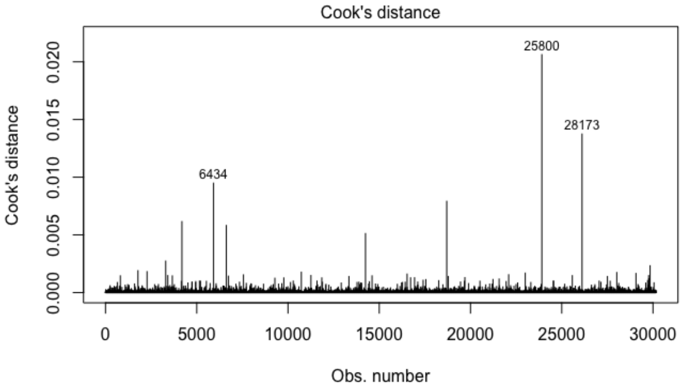

Figure 3: *Cook's distance of the Final model*

Finally, we evaluate the models by computing their accuracy and their AUC score. The accuracy corresponds to the number of true positive and true negative divided by the total amount of predictions. The ROC curve (Receiver Operating Characteristics) is a plot of the True Positive Rate $\frac{TruePositif}{TruePositif+FalseNegatif}$ against the True Negative Rate $\frac{FalseNegatif}{TrueNegatif+FalsePositif}$. From this, we compute the AUC (Area Under the Curve) score which is a good metric taking both type of error into account (False Positive error and False Negative error). Note however that these metrics have the drawback of being sensitive to class imbalances which is the case in our dataset.

The Final model has an accuracy of 84.63% and an AUC score of 0.900, whereas

Kohavi's model has an accuracy of 84.73% and an AUC score of 0.902. Hence the Final model performs nearly as well as Kohavi's model even if it uses much less variables. The ROC curves of both model are shown in Figure 4.

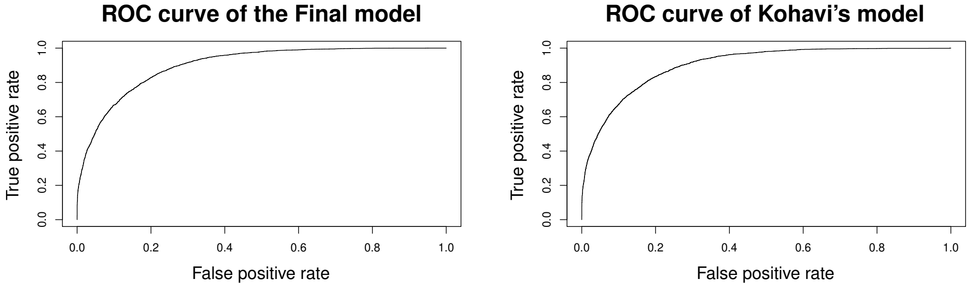

Figure 4: ROC curves of the models

5 Final model

The Final model is as follow

$$\pi(x) = P(Y = 1|x) = \frac{exp(\beta_0 + \sum_{i=1}^{8} \beta_i x_i)}{1 + exp(\beta_0 + \sum_{i=1}^{8} \beta_i x_i)}$$

where x_1 is the age, x_2 is the sex, x_3 is the education number, x_4 is the relationship, x_5 is the occupation, x_6 is the capital gain, x_7 is the capital loss and x_8 is the hours per week. Note that x_4 and x_5 are both categorical variables and are in fact sets of indicator variables representing each factor. The details of the model are written in Table 3 in appendix.

6 Conclusion

In conclusion, we have shown that it is possible to predict the income level of a population using a logistic model based on socio-demographic variables. Whereas Kohavi used 14 variables in his model, it has been shown that only 8 variables mattered to have a model that performs equally well.

Our final logistic model has an accuracy of 84.73% and an AUC score of 0.900. Even if thoses results are good, it is important to point out that there is a class imbalance in the dataset used (both in the training and the test data) which may alter the model.

Also the diagnostics performed on the Final model have shown that there is not a linear relation between the logit and some of the predictors of the model which indicates that a logistic model might not be the most suited approach for this problem.

Finally, we can reject the null hypothesis that none of the socio-demographic variables used in this study have a statistically significant impact in predicting whether a person makes over 50'000$. Indeed, the age but most importantly being a male is correlated with a higher income which does not respect the Title VII of the Civil Rights act of 1964. This analysis should be repeated using more recent data from the Census Bureau Database to see if the situation has evolved in the US since 1994.

7 Appendix

Table 3: *Summary of the Final model*

	Estimate	Std. Error	z value	Pr(> $\|z\|$)
(Intercept)	-6.84	0.18	-38.60	< 2e-16
age	2.76e-02	1.59e-03	17.36	< 2e-16
sex Male	0.83	7.84e-02	10.64	< 2e-16
education_num	0.28	9.42e-03	29.57	< 2e-16
relationship Not-in-family	-1.87	5.71e-02	-32.80	< 2e-16
relationship Other-relative	-2.03	0.20	-10.14	< 2e-16
relationship Own-child	-2.99	0.14	-20.85	< 2e-16
relationship Unmarried	-1.86	9.89e-02	-18.82	< 2e-16
relationship Wife	1.27	0.10	12.29	< 2e-16
occupation Armed-Forces	-0.59	1.48	-0.40	0.69
occupation Craft-repair	-1.69e-02	7.84e-02	-0.22	0.83
occupation Exec-managerial	0.74	7.52e-02	9.89	< 2e-16
occupation Farming-fishing	-1.33	0.14	-9.73	< 2e-16
occupation Handlers-cleaners	-0.74	0.14	-5.18	2.21e-07
occupation Machine-op-inspct	-0.32	0.10	-3.16	1.55e-03
occupation Other-service	-0.95	0.12	-8.09	6.21e-16
occupation Priv-house-serv	-3.72	1.75	-2.12	3.38e-02
occupation Prof-specialty	0.45	7.81e-02	5.81	6.40e-09
occupation Protective-serv	0.43	0.12	3.58	3.46e-04
occupation Sales	0.22	7.99e-02	2.72	6.55e-03
occupation Tech-support	0.61	0.11	5.57	2.56e-08
occupation Transport-moving	-0.17	9.79e-02	-1.71	8.70e-02
capital_gain	3.16e-04	1.05e-05	30.13	< 2e-16
capital_loss	6.37e-04	3.81e-05	16.73	< 2e-16
hours_per_week	3.01e-02	1.66e-03	18.14	< 2e-16

References

[1] Alboukadel Kassambara. *Logistic Regression Assumptions and Diagnostics in R.* 2018. URL: `http://www.sthda.com/english/articles/36-classification-methods-essentials/148-logistic-regression-assumptions-and-diagnostics-in-r/` (visited on 07/06/2022).

[2] Ron Kohavi. "Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid". In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining.* KDD'96. AAAI Press, 1996, pp. 202–207.

[3] Zach. *How to Identify Influential Data Points Using Cook's Distance.* 2019. URL: `https://www.statology.org/how-to-identify-influential-data-points-using-cooks-distance/` (visited on 07/06/2022).

[4] Alexandre Zajic. *Introduction to AIC — Akaike Information Criterion.* 2019. URL: `https://towardsdatascience.com/introduction-to-aic-akaike-information-criterion-9c9ba1c96ced` (visited on 07/06/2022).