MA429 Formative Group Project

Group 10

33769

25259

A report submitted to the Department of Mathematics of the London School of Economics and Political Science

March 12, 2024

Abstract

This report delves into the analysis of the UCI Adult dataset [1], derived from the 1994 US Census, encompassing variables like occupation, age, and educational attainment. Our objective is to demonstrate the significance of these variables in predicting an individual's income. Our findings reveal that several variables have minimal impact on income prediction, as evidenced by the sparse non-zero coefficients in the Lasso regression and the shallow nature of the decision trees. Nonetheless, *age*, *hours worked per week*, and *years of education* emerge as the most influential factors.

In our exploration of four models — Decision Trees, Bagging, Random Forests, and Lasso Regression — we discover that the Random Forest model exhibits superior performance. Although our model achieves an accuracy of 83.6%, slightly below the benchmark of 85.2%, it excels in precision, attaining a score of 91.8%, well above the benchmarked 80.2%. This highlights the model's effectiveness in identifying individuals with higher incomes accurately, despite some deviations in overall accuracy.

Despite achieving notable accuracy and precision, our study faces limitations, particularly regarding the dataset's age. This raises questions about the model's applicability to current socioeconomic conditions. The data, being two decades old, may not accurately reflect today's economic and social dynamics, thereby limiting its relevance for contemporary income prediction. However, when focusing on the socioeconomic context of the 1990s, our model provides valuable insights.

Looking forward, future research should aim to incorporate more recent data, experiment with additional machine learning models, and conduct comprehensive performance comparisons. Expanding the dataset and exploring innovative modeling techniques could enhance our understanding of income determinants and improve prediction accuracy. This endeavor would not only address the limitations of our current study but also pave the way for more nuanced and applicable income prediction models in the future.

Contents

Chapter 1

Introduction

In this chapter, we discuss our exploratory data analysis and any pre-processing we performed. Understanding the data will be a crucial part of good modelling, so transforming the variables correctly is of high importance. Furthermore, we consider the effects of feature engineering, in particular the handling of missing values, dropping of "irrelevant" variables, and dealing with weighted data.

1.1 Dataset

The dataset is a census data conducted in 1994, obtained from UC Irvine Machine Learning Repository. The objective of collecting the dataset is to predict whether a person's income exceeds $50,000 per year. This dataset contains of 48,842 instances with 14 features. We summarise the data in Table 1.1.

1.2 Exploratory Data Analysis

Data pre-processing is conducted prior to modelling. Initially, we labeled the features and converted the categorical data types into levels using *as.factor()* to facilitate analysis during modeling. Moreover, we identified missing values in three features: *workclass*, *occupation*, and *nativecountry*.

We first addressed the missing values in *nativecountry*. This feature contains 42 unique values, with 89.59% of the sample identifying the United States as their native country and 1.79% of the data missing. Identifying the reasons behind the missing *nativecountry* values is challenging, as these do not appear to be related to other features. Additionally, the missing data in this feature are nearly independent of the missing values in the other features. As such, we decided to do nothing for the missing values, with further adjustments made in feature engineering section.

We then addressed the missing values in the *workclass* and *occupation* attributes. Upon examination, we found that 5.64% of the entries for *workclass* and 5.66% of

Variable	Data Type	Number of Levels	Missing Values
age	Numerical	Integer	n/a
workclass	Categorical	9	5.64%
fnlwgt	Numerical	Integer	n/a
education	Categorical	16	n/a
educationnum	Numerical	Integer	n/a
maritalstatus	Categorical	7	n/a
occupation	Categorical	15	5.66%
relationship	Categorical	6	n/a
race	Categorical	5	n/a
sex	Categorical	2	n/a
capitalgain	Numerical	Integer	n/a
capitalloss	Numerical	Integer	n/a
hoursperweek	Numerical	Integer	n/a
nativecountry	Categorical	42	1.79%
income	Categorical	2	n/a

Table 1.1: A summary of the variables.

the entries for *occupation* were missing. Interestingly, 5.64% of the dataset had simultaneous missing values in both *workclass* and *occupation*. Additionally, 5.66% of the dataset had at least one missing value in either of these features. Our preliminary hypothesis suggests that respondents hesitant to disclose their occupation might also be reluctant to specify their work class, and vice versa. This reluctance could stem from the nature of their work, which perhaps makes them reticent. Supporting this hypothesis, further analysis revealed that 89.60% of the cases with missing values in both features reported incomes below $50,000. Consequently, we opted to retain these missing values as is, segregating them from the complete entries for separate analysis.

Finally, in addition to addressing missing values, we converted the income categories into a binary variable, where TRUE indicates an income above $50,000 (denoted as '>50K') and FALSE denotes an income of $50,000 or less (denoted as '$\leq$50K').

1.3 Feature Engineering

Several adjustments have been made to the features: *nativecountry*, *education*, *educationnum*, *capitalgain*, and *capitalloss*.

The *nativecountry* feature contains 42 unique values. Given the limitation of

modeling a decision tree in R, which allows a maximum of 32 levels, we transformed this feature into binary variables: TRUE represents samples whose native country is the United States, and FALSE represents samples from other countries.

Furthermore, *education* and *educationnum* convey the same information, namely the level or years of education attained by the sample. The difference is that education is categorical, whereas *educationnum* is numerical, with values ranging from 1 to 16, which corresponds to the 16 levels in *education*. Based on this redundancy, we decided to remove *education* to avoid duplicating information.

Lastly, *capitalgain* and *capitalloss* both relate to financial outcomes from capital investments, albeit from opposite perspectives. We combined these features by incorporating *capitalloss* into *capitalgain*, converting *capitalloss* values to negative when merging them into *capitalgain*. However, after observing that 87.01% of the values in *capitalgain* are zero, we hypothesised that this might be due to respondents' reluctance to report capital gains (possibly to avoid taxes), hesitance to disclose capital losses (perhaps due to embarrassment), or simply because they had not engaged in investments—an assumption that may seem unlikely unless the respondents lack financial literacy or experience. Consequently, we decided to remove this feature from our analysis.

Chapter 2

Modelling

2.1 Lasso Regression

For our initial model, we used a Lasso regression. We did this for two reasons: firstly, Lasso is one of the simplest models so it provides a baseline performance to compare our future models against. Secondly, it will also be able to inform us on the amount of regularisation required in other models. Beyond the performance statistics, we are very interested in how the amount of regularisation affects the magnitude of the coefficients. We used the glmnet package [2] to leverage the built-in methods to analyse performance.

In Figures 2.1 and 2.2, we are able to learn more about the learning process and the effect of regularisation on performance. We should point out on Figure 2.1 the "top" axis $(4, 4, 4, 3, 3, 1)$, which represents the number of non-zero coefficients at the different values of $log(\lambda)$. Recall that our dataset contains 14 features (before feature engineering) so even with no regularisation, Lasso sets many variables to zero.

Further, Figure 2.2 shows shows a 10-fold cross-validation, showing the mean and standard deviation at different values of lambda. The two dashed lines at approximately $log(\lambda) = -7.25$ and -4 correspond to the values lambda that gives the minimum error rate, and 1 standard error above the minimum across the 10-folds, respectively. The latter corresponds to a more conservative estimate to prevent overfitting the final model. These values of $log(\lambda)$ solve to give $\lambda_{min} = 6.6 \times 10^{-4}$ and $\lambda_{1se} = 0.023$.

2.2 Decision Tree

We then decided to use a decision tree. This decision was based primarily on two reasons: First, decision trees are quite straightforward and easier to interpret compared to complex models like neural networks. Additionally, they have the capability to handle both numerical and categorical features without the need for comprehensive

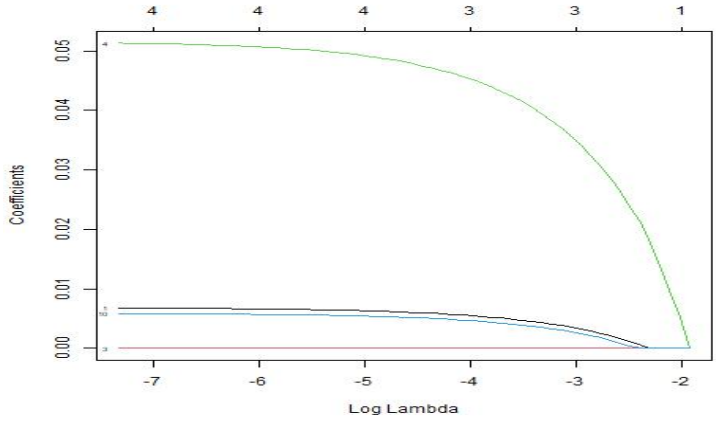

Figure 2.1: Our Lasso model coefficient values against $log(\lambda)$.

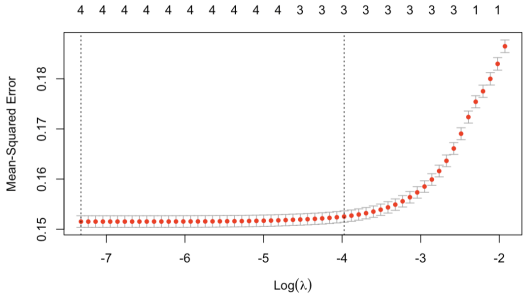

Figure 2.2: Cross validation, training error against $log(\lambda)$

preprocessing. We also explored weighted decision trees, a variation that incorporates *fnlwgt* as our weight during the training process, non-weighted ones and pruning. However, the weighted decision tree is preferable in this case, mainly because the *fnlwgt* data allows us to prioritise observations with a higher tally in the dataset, indicating a higher probability of occurrence. At the same time, the pruned tree gave the same output like the weighted decision tree. We visualise the decision tree in Figure 2.3 using the rpart package [3].

As seen in Figure 2.3, only five features are utilised: *relationship*, *educationnum*, *occupation*, *age*, and *hoursperweek*, with an addition of *fnlwgt* as the weight. The tree comprises a total of 11 nodes and 10 branches, including 6 leaf nodes and 5 decision nodes. It is important to note that from the 21,503 instances (70% of the dataset) in our training set, 75.28% are labeled as FALSE (income of $50,000 or less), and 24.72% are TRUE (income above $50,000). Among them, 55% of the training set, characterised by a *relationship* of either *not-in-family*, *other-relative*, *own-child*, or *unmarried*, predominantly earns $50,000 or less.

For those not classified in the aforementioned category with an *educationnum* of at least 13 — which applies to 14% of the training set — the majority earn an income

above $50,000. If their *educationnum* is less than 13 and their *occupation* falls into one of the following categories: *armed-forces*, *craft-repair*, *farming-fishing*, *handlers-cleaners*, *machine-op-inspct*, *other-service*, *priv-house-serv*, or *transport-moving*, they are highly likely to have an income of $50,000 or less, unless their *hoursperweek* is at least 34. For those working at least 34 *hoursperweek* — representing 9% of the training set — there is a likelihood of earning an income above $50,000.

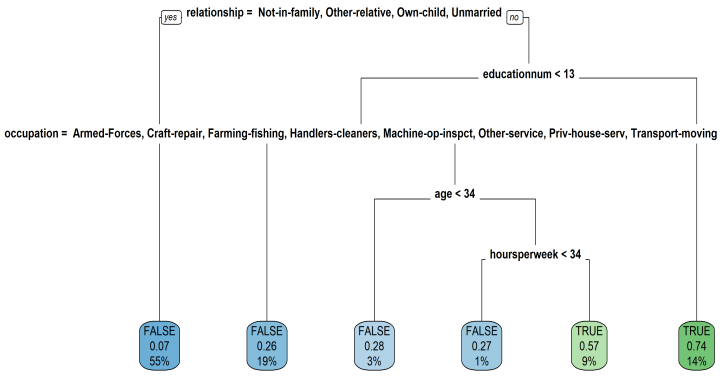

Figure 2.3: Our final Decision Tree

In simple terms, this suggests that individuals with a higher education level of at least a bachelor's degree (an *educationnum* of 13 corresponds to a Bachelor's degree in *education* according to the dataset) are highly likely to earn above $50,000. However, even for individuals whose education level is below a Bachelor's degree, the likelihood of earning above $50,000 increases if they work at least 34 hours per week.

2.3 Random Forests

Building upon the decision tree results, we attempted a Random Forest model to improve the 82.5% accuracy which is slighlty worse than we anticipated. We tuned the model by comparing the performance as we changed the number of randomly-selected predictors allowed at each split. We repeated this for $p = 1, 2, \ldots, 10$ predictors, and we summarise our findings in the Table 2.1 below.

We see that $p = 2$ achieves the highest accuracy of 83.6%, and so choose this one as our "best" forest model. We believe that accuracy is the most important metric for this task, however we are conscious of the fact that other metrics may be more valuable to optimise with different objectives. Moving forward, however, we

p	Accuracy	Precision	Recall
1	82.7	83.6	95.6
2	83.6	87.1	91.8
3	83.3	87.2	91.0
4	82.6	86.9	90.5
5	82.4	87.0	90.0
6	82.0	86.7	89.7
7	82.0	86.8	89.6
8	82.1	86.8	89.6
9	82.2	86.9	89.7
10	82.2	87.1	89.5

Table 2.1: A performance summary of the random forest models.

the random forest model refers to $p = 2$, such as in Table 3.1.

2.4 Bagging

We opted to investigate bagging as a potential means to enhance our model's performance. As depicted in Figure 3.1, the results were not particularly impressive (importantly, worse than the decision tree and random forest in all 3 metrics), leading us to conclude that further emphasis on this model is unwarranted.

Chapter 3

Conclusion

3.1 Findings

In Table 3.1 we briefly summarise the models we used.

	Model Complexity	Regularisation	Accuracy	Precision	Recall
Lasso	Simple	Low	78.0	75.3	79.7
Decision Tree	None	Low	82.5	87.6	89.3
Random Forest	Complex	High	83.8	87.2	91.8
Bagging	Complex	Low	82.1	87	89.1

Table 3.1: A performance summary of the models we used.

3.2 Limitations

There are a few key limitations to this study that we need to consider. First, using census data from 1994 might not give us a clear picture of today's economy or society. A lot has changed since then, which could make our predictions less accurate for the present day. Second, there's a chance that important information was missed when collecting the census data. Back in 1994, collecting data wasn't as easy as it is now because technology wasn't as advanced. This missing information could weaken our analysis. Third, there's a concern that our model might be too focused on the specific patterns found in the old data, meaning it could be really good at predicting the past but not as good when it comes to new or current information.

While there might be other smaller issues, the biggest thing to keep in mind is whether or not old data like this can help us understand the present. If we plan to use this model for today's world, we might need to look into getting more recent data to improve its accuracy.

3.3 Further Work

In the future, several enhancements could be considered to improve the research. Firstly, utilising more recent census data or conducting a new census that accurately mirrors current socioeconomic conditions would be beneficial. Secondly, it's important to re-evaluate and potentially expand the variables or features included in the analysis to gain a more comprehensive understanding of factors influencing income today. Relevant additions might include the name and ranking of one's educational institution, their level of technological proficiency, the industry of employment, and any criminal records. Including geographical factors is also crucial, as location significantly affects income distribution—for example, salaries in London are generally higher than in Manchester for comparable positions. Lastly, adopting more sophisticated modeling techniques could help minimise the risk of overfitting and enhance the accuracy of predictions.

Bibliography

[1] Becker, B. and R. Kohavi (1996). Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

[2] Friedman, J. H., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 33(1), 1–22. URL https://www.jstatsoft.org/index.php/jss/article/view/v033i01.

[3] Therneau, T., B. Atkinson, and B. Ripley (2021). rpart: Recursive partitioning and regression trees URL https://CRAN.R-project.org/package=rpart, r package version 4.1-15.