Liutauras Kuolis s1060854
Dinesh Kumar s1050577

Feature Importance Analysis of the 1994 US Adult Income Dataset

Abstract

In this paper, we attempt to use US Adult Census data to discover the most significant features/attributes that determine the income level of an individual (whether they earn more or less than 50k USD) by using different classification algorithms. From results we found that sex is a big factor in choosing the individuals income. Although different models exhibited different results.

Jupyter file attached with the report.

Introduction to Problem and Rationale

Trying to predict a person's financial wellbeing is quite a tricky task and is mainly dependent on the relationship between their income and expenditure, which consequently affects their savings.

Expenditure depends on the number of responsibilities one takes on, but how much they choose to allocate to each of these responsibilities still remains under their control. However, income may be affected by multiple factors that are out of an individual's hands.

For example, theoretically the value of an individual should be solely based on their skills/merit, and it is often the case that people who have the supporting education and experience are more enticing candidates as they're able to back up the skills they claim to possess, however other factors like their sex, race, family status and age can also affect their income.

Individually, this exploration will help us identify which of these attributes affected an individual's income the most in 1994 – this can be useful for people to determine the factors in their control that they can change for better financial prospects and to be able to advantageously position oneself in an increasingly competitive world.

On a societal level, analyzing the data will also give us insight into whether there existed income inequality by checking if factors outside one's control (such as those mentioned above) significantly impact one's financial wellbeing, and then trying to figure out how we can deal with the bias (as in an ideal society we wouldn't want that income inequality existed.)

This is also useful for the future, because as we're being ushered into this new era, traditional definitions of education and skill are changing, where instantly discarding work candidates without the 'acceptable' years of work experience and education level is not very wise as learning goes digital and experience comes from non-traditional means such as personal projects, and thus for the attributes that are most significant determining what exactly makes then significant and aligning them with the advancements that have taken place since the collection of the data.

Liutauras Kuolis s1060854
Dinesh Kumar s1050577

Keywords: supervised learning, binary classification

Related Works

We have found three papers that have been published before. All three of them go more in-depth in this subject. The feature significance analysis in these papers is focused more on the specific values of the attributes, rather than in attribute as a whole.

1. Feature Significance Analysis of the US Adult Income Dataset by Junda Chen
2. Predicting if income exceeds $50,000 per year based on 1994 US Census Data with Simple Classification Techniques by Chet Lemon, Chris Zelazo and Kesav Mulakalur
3. Supervised Learning for Binary Classification on US Adult Income by Li-Pang Chen

Methodology

Dataset

To investigate this, we found the US Adult Census Income Dataset most appropriate because it is quite a popular dataset for which finding supporting research would be relatively easier given this is our first machine learning project, and also because there are plenty of attributes as well as data instances, which provides quite a lot of information for the algorithms were going to use.

Attributes

The US Adult income dataset was extracted from the 1994 US Census Database and consists of 48,842 different records with anonymous social information on the following 14 attributes:

Index	Feature Name	Description	Data type
1	age	-	Continuous interval
2	workclass	the employment status of an individual in relation to their pay and employment history	Discrete categorical nominal
3	fnlwgt	Final weight/ the number of other people in the dataset represented by this person	Continuous
4	education	Education level	Discrete categorical ordinal
5	education-num	Years of education acquired	Continuous interval
6	marital-status	-	Discrete categorical nominal
7	occupation	-	Discrete categorical nominal
8	relationship	relationship in terms of the family	Discrete categorical nominal
9	race		Discrete categorical nominal
10	sex	Gender assigned at birth	Discrete categorical nominal
11	capital-gain	Capital gain in US dollars	Continuous interval
12	capital-loss	Capital loss in US dollars	Continuous interval
13	hours-per-week	Hours worked per week	Continuous interval
14	native-country	Country of birth	Discrete categorical nominal

Source: https://archive.ics.uci.edu/ml/datasets/adult

These are then accompanied by a final binary class label y (a Boolean) representing whether or not a particular individual earns greater than or equal to 50k usd and this would be determined using its attributes.

Preprocessing Methods

Missing values

Missing values in this dataset were being represented by a '?', and upon inspection, the total number of rows that have these total to 3620. Since this is a relatively insignificant number of 'bad' entries (compared to the total number of rows in the dataset), and because the number of missing values in the 'work-class' and 'occupation' attribute seem linked due to the similar number of missing values, we thought the best way to deal with those would be to just remove them instead of trying to fit it to a general average.

Outliers

Upon inspecting the data distribution using a pandas built in function describe(), some seemingly unusual values were present – namely the maximum capital gain at 99999 and the maximum hours worked per week at 99. Upon further inspection, the weekly hours apparently are not very unusual for American standards as there were other records that had similar hours, however for capital gains, our intuition was right as it doesn't make a lot of sense as when there's no data between 50-98k except the 229 values at 99k, so those instead of being removed were instead replaced by average capital gains as they likely stem from an input error

Attribute simplification

- Removed rows which have work class *Never-worked* and *Without-pay*

This was done because these would fall on the lower extreme of the income bracket, as people without pay or those that have never worked would receive little to no income, which means such records would serve as outliers and may affect the accuracy of the classification.

- Also combined *federal*, *state* and *local gov* to just **government**
- Combined *1st-4th* and *5th-6th* to **primary school**
- Combined *7th-8th* and *9th* to **middle school**
- Combined *10th*, *11th*, *12th*, and *HS-grad* to **high school**
- Combined *Assoc-acdm, Assoc-voc, Some-college* and *Prof-school* to **associates degree**

Such simplification was performed to reduce unnecessary complexity of the data which may affect classification attempts in unintended ways, and combining and grouping similar attributes this way should help achieve that effect.

- Normalization of numerical values

This helps make sure that differences between numerical values are standardized so that values that values are not interpreted being as different as they appear to be, and are interpreted in the context of other values of the same attribute/column.

- Encoding nominal attributes to be represented by numbers using Label Encoder (we encoded the attributes to make our CART and Logistic Regression algorithms could work properly, since both models can't take objects as an input)

Done later for comparison:

- Removed *fnlwght* and *relationship in family* and used *education-num* instead of *education* column.

We didn't really understand the purpose of final weight, and being unsure of how it would affect the classification, we decided to not include it.

We removed relationship in family because the combination of relationship and gender seems to provide the same information and thus its just an extra unnecessary attribute.

Similarly, we decided instead of trying to group together schooling to just get rid of the education column entirely and use education num column instead as it summarizes the same information.

Approach

We decided to use classification algorithms CART Decision Tree and Logistic Regression. The main reason responsible for this decision was that both of these algorithms are for classifying and they use completely different strategies. While our CART algorithm uses GINI index to create the best set of splits, Logistic Regression on the other hand uses likelihood.

Results and Analysis

Decision Trees (CART)

For us this model worked better than logistic regression. The accuracy of our decision tree algorithm is 85,4% for the training set and 85,2% for our test cases which is relatively good. The binary data here worked perfectly. Figure 1 shows the significance of all of our attributes. The value that is shown in Figure 1 is in percent, if the bar is higher, it means the significance is stronger.

Some of the features were unexpected at first, but when we did more research, we realized what was happening. Relationship in family is the most significant attribute, because it has a value called

husband. In the first step of our decision tree model husband is filtered out from everything else like wife, not-in-family, own-child and more.

The attributes education years and capital gain make sense to be really significant, since that usually shows individuals' skills.

Other attributes weren't that significant. For example, race and sex we thought would matter more. The explanation might be that for sex our model instead just used relationship in family, since that shows the sex and marital status (someone who is male and a husband will probably earn more than someone who is male, but not a husband). Race has a lot of values which don't have that many people representing so that might have influenced the decision tree.

Figure 3 shows the results after we removed some values. We removed 2 values that didn't have much to do with our algorithms like final weight and education (we have education years). Additionally, we removed relationship in family value, because we wanted to see how that would change the accuracy of our algorithms and importance.

As we can see from Figure 3 removing mentioned attributes had a big impact on marital status. Marital status is now the most important feature of this dataset. Moreover, the accuracy of our model improved by around half percent. In addition, education years and capital gain importance grew a little too. The change from relationship in family to marital status can be explained. We think it is because the marital status is more important for our model than sex, but when we have both of those combined into one (relationship in family), the algorithm chooses it instead.

Logistic regression

This algorithm was less successful, its accuracy was 82% for both testing and training. In Figure 2 we can see the importance of each feature. The absolute value in Figure 2 shows how important the feature was. The sign in these values represent whether the feature is positively related to the label or not.

From what we can see education years, capital gain two of the most important features for this model, the same as in CART, but the third one and most important is sex. This might be because if the attribute is not binary or continuous the model it is underestimated. We can see that education years, capital gain and sex worked really well while attributes like marital status, occupation, education and other nominal attributes got undermined.

Figure 4 shows the feature analysis with the same attributes removed as in Figure 3 just to see how it impacts the performance of the algorithm. It didn't have a significant change in accuracy of this model. Although, the feature importance has changed. Every feature's importance got amplified a

little, except for sex which got amplified a lot more than other attributes.

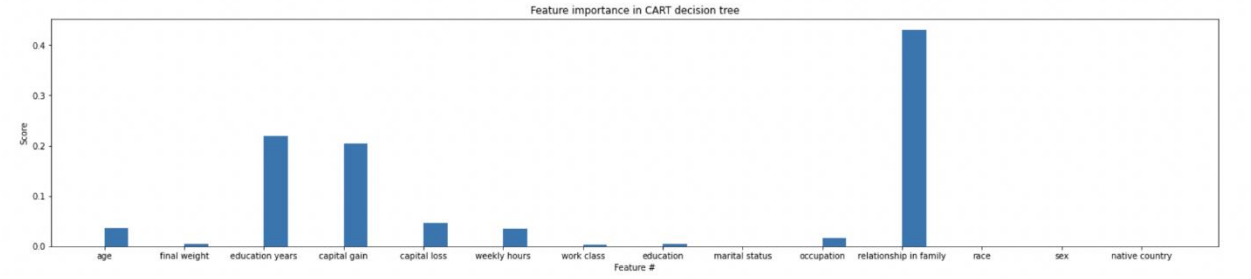

Figure 1: This figure shows how each feature influences our model choice in percent. Features 10, 2 and 3 has the highest importance

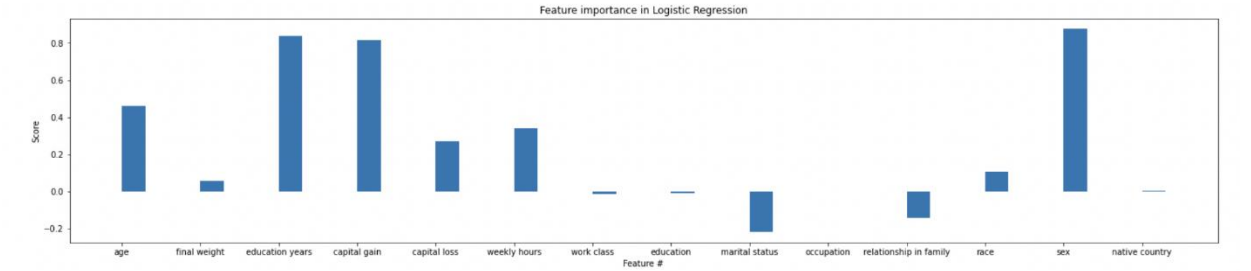

Figure 2: This figure shows how each feature influences our model choice. Features 12, 2 and 3 has the highest importance

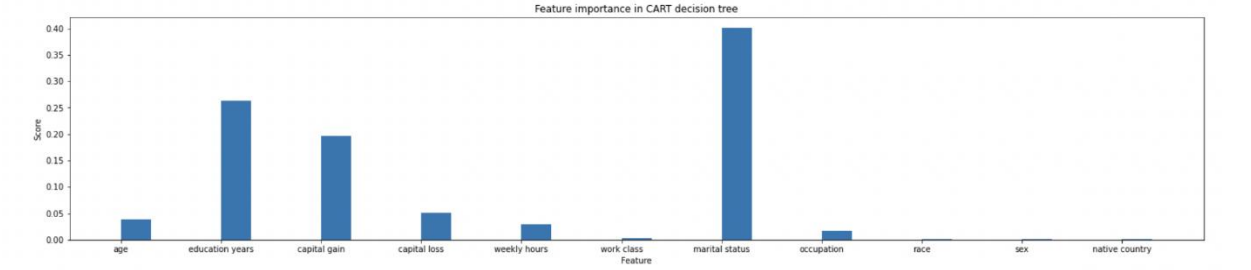

Figure 3: This figure shows how each feature influences our model choice in percent. Features 1, 7 and 10 were removed

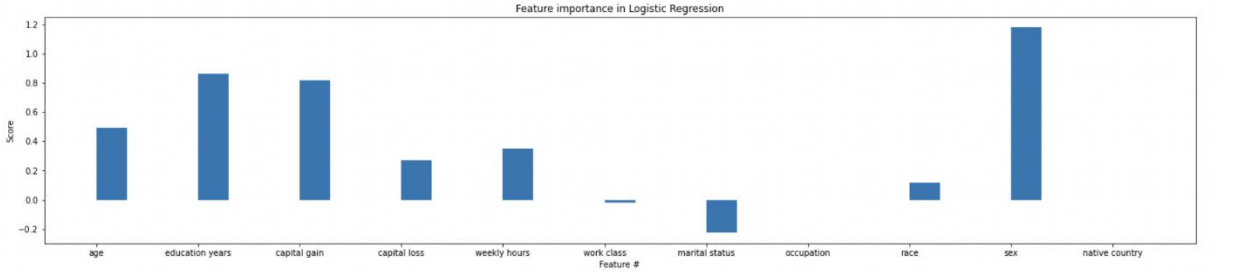

Figure 4: This figure shows how each feature influences our model choice. Features 1, 7 and 10 were removed

Evaluation

From what we done here we are pretty happy with our results. We should have done a better job. We just removed the entries with missing values instead of adapting to them. However, this should not have made a big difference since we removed only fraction of the entries.

We also think we should have done more research on the actual values as the mentioned papers in Related Works did. The outcome was mostly what we expected. However, we thought that age would influence our models more than it did, this might be because we had a feature which showed

education years and that was more interesting attribute to our models than age. We never thought that relationship in family would influence the model as much as it did. Although, we expected sex to be a big factor, but relationship in family swapped out sex in our decision tree model.

For the train, test results we thought that Logistic Regression would do worse than CART, since there are a lot of nominal attributes which are not that great for Logistic Regression. Our hypothesis was true, the train/test accuracy differed by about 3%.

All in all, we think we did decently in this project. Although, we still have some paths to follow up with this research.

Further Research

Possible future directions would be putting more effort into cleaning up the data and trying out new methods for training and testing the data. Additionally, we should do a more in-depth analysis of the feature significance with actual attribute values instead of using the feature as a whole. Then we could follow-up by doing the same thing on a more recent data set to see how it has changed over the years.

References

[1] Adult Data Set **https://archive.ics.uci.edu/ml/datasets/adult**

[2] Kaggle **https://www.kaggle.com/johnolafenwa/us-census-data**

[3] Feature Significance Analysis of the US Adult Income Dataset by Junda Chen

[4] Predicting if income exceeds $50,000 per year based on 1994 US Census Data with Simple Classification Techniques by Chet Lemon, Chris Zelazo and Kesav Mulakalur

[5] Chen L-P. Using machine learning algorithms on prediction of stock price.

[6] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. Foundations of machine learning.