Predict Salary Income using Machine Learning

Submitted by

Bhagyashree Chavan

Gulkhan Anassova

Trineth Reddy

Dept: Data science and business analytics

Professor: Dr. H. Wang

February 28th, 2021

Saint Peter's University Jersey City, NJ 07306 United States of America

ABSTRACT

The project objectives to apply machine learning models which would be the best fit model to the problem of identifying or predicting the salary income of an individuals is greater than 50k. The data sourced by UCI machine learning repository called ‘Adult dataset’. The dataset contains demographic information such as age, level of education and occupation and employment type etc. The analysis process approach includes data pre-processing, categorical feature representation, data imbalance handling, modelling, evaluation of datasets and normalization. In this process topics will be covering such as statistical analysis, features reduction, missing values, elimination of special characters, exploratory analysis, different Machine learning models. Based on analysis values of each attribute we try to predict the income of an individual.

Keyword: Decision Tree, Random Forest, Ada boost, Predictive Analysis, Machine Learning, and Grid Search.

I. INTRODUCTION

Adult dataset is fairly old, and it was extracted from the 1994 Census bureau database of United States of Government. The dataset was taken from UCI Machine Learning repository and Extraction was done by Barry Becker from the 1994 Census database. This dataset has information about the annual income of individuals from 42 different countries but 90% of data dominated by the United States and rest of the data contains 8% of Mexico category and 2% other 40 countries.

The dataset has different columns with categories based on age, work class, education, numerical form of education, marital status, occupation, relationship, race, gender, etc. We have 48842 sample set and 15 different variables.

The dataset has many attributes but here considering following is most important variables to get the income of an adult based on the data available for this classification.

1. Age: Discrete- The individuals age from 17 to 90

2. Work class- Private, Federal-Government, etc - Nominal it has 9 categories

3. Education level of education- Ordinal- It has 16 categories

4. Marital Status: Nominal- It has 7 categories

5. Occupation- Nominal It has 15 categories

6. Race- Nominal- It has 5 categories

7. Hours individual work per week- Discrete- It has 1 to 99 hrs

8. Income (whether or not an individual makes more than $50,000 annually): Boolean (≤$50k, >$50k)

II. METHODOLOGY

The dataset is obtained from UCI ML repository and it may contain disturbance for the data analysis, it is required for the data to be maintained such that there would be no errors during the analysis. Hence, several data cleaning techniques have been employed as required for the acquired dataset.

Data Pre-processing

Data pre-processing is the process of cleaning and preparing the text for further analysis. Online datasets, usually, may contain lots of noise and uninformative parts that could hinder the output during the analysis. To keep all the datasets in line for statistical analysis, following data cleaning methods have been employed to get the best clean data to improve the analysis. This would essentially decrease the noise in the datasets and would allow us to obtain more precise statistical values. After we import data, we check the head of the data frame, all the columns available, shape of the data and obtain the statistical summary to better understand our dataset. Once we finish this step, we can check for the data cleaning process.

Filling Missing Values

In this step we check for the missing values in each attribute of the dataset or the value is n/a. If any we fill it with 0 or median depending on the null values. As there are no null values, we proceed to next step of data cleaning.

Finding Special Characters

Any special characters in data could hinder the further steps when we convert the categorical variable to numerical values. So, we have decided to check for the special characters in the data and we found work class, occupation and native country had special characters. We have decided to remove all the rows containing the special character with 'nan' and remove all the rows containing 'nan'. After the above step we checked if there are any special characters, and no special characters were found in the dataset. The dataset was reduced by 3620 rows and the shape of the dataset is now (45222, 15).

Feature Engineering:

With few exceptions, Machine Learning algorithms do not perform well when the input numerical attributes have very different scales. Machine Learning model requires input data in numerical notations to extract patterns from it and make predictions. But not all the data provided in our source dataset is numerical.

Mapping Categories to Numerical

We have converted categorical variables to numerical variables. Now gender, race, marital status, work class, education, occupation, and relationship are numerical variables. Mapping the data into numerical data using map function.

Selection of Model and Normalize data:

The data is split into 70% for the training and 30% for the testing. We have used StandardScaler() for standardizing and train_test_split to split the data using Scikit-learn library.

Models we are considering Decision tree which is used for supervised learning techniques and it starts with all training samples as root node to top of subsequent layer which will provide maximum information. The function has been used DecisionTreeClassifier()

Another model is Random forest model which is also use for supervised learning algorithm it provides good results without hyperparameter tuning most of the time. The function we have used RandomForestClassifier() to perform analysis.

III. **ANALYSIS**

Exploratory Data Analysis (EDA)

Race Distribution

The Maximum percent of the earnings in the dataset are Whites with 86.03 percent second to Blacks.

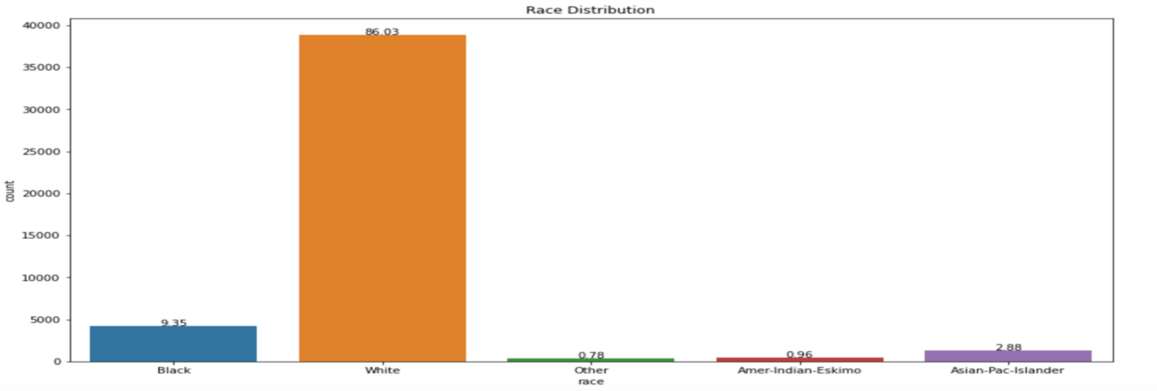


Figure 1: Race Distribution

Work Class vs Income

We check the working-class distribution against the annual income of the person, and we could see that most Self-Employed people make less than 50k next to Federal Government employees and Local Government employees are the highest who make more than 50k annually. A Very few individuals makes more than 50k annually in private sector.

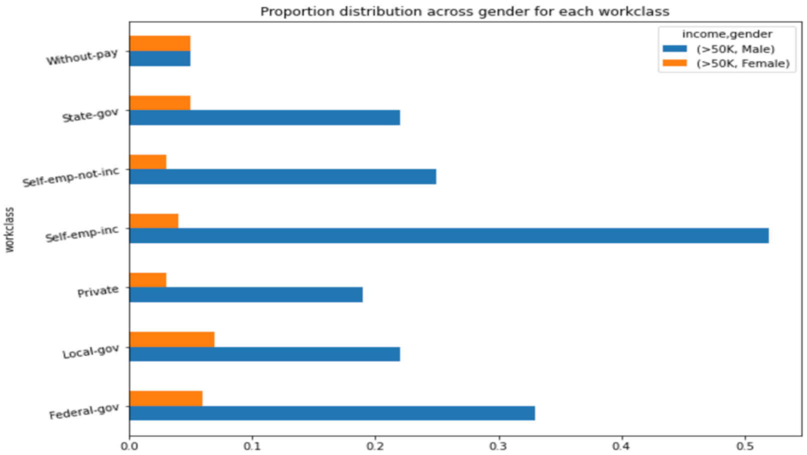


Figure 2: work class vs Income

Occupation vs Population proportion

We visualized the proportion of population between the two income segments to understand which occupation has the most income difference in the number of people working. The interesting findings of this visualization is that the Private house service has most difference between income and almost no one makes over $50k whereas, Executive managerial has minimal difference between the two income groups.

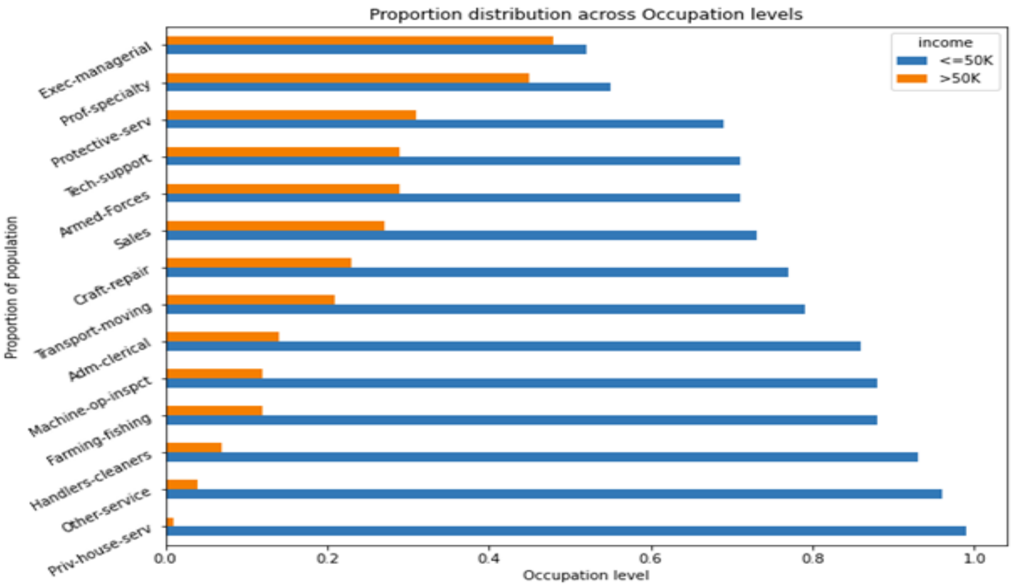


Figure 3: Occupation vs Population proportion

IV. FEATURE ENGINEERING

In this part of process, after using many models these five classifiers over perform other models which is mainly used for analysis: Random forest, Decision Tree, Random forest with Grid search, Decision tree with Grid search and Adaboost. Before processing data provided are Categorical data like Work Class, Education, Occupation, etc. we need to convert these into numerical notations.

Running a loop value_counts of each column to find out unique values. some of the data provided are unique like the ‘workclass’ attribute which has only 7 distinct values and some columns have a lot of distinct values like ‘fnlgwt’ attribute which has around 2000+ values. So, let us drop the attributes that have noisy data.

We have dropped the noisy data with columns 'educational-num', 'hours-per-week', 'fnlwgt', 'capital-gain', 'capital-loss', 'native-country' and we are left with the following columns in dataset 'age', 'workclass', 'education', 'marital-status', 'occupation' etc.

Training and Evaluating on the Training Set

TYPES OF MODEL	ACCURACY
RANDOM FOREST	81.462%
DECISION TREE	81.432%
RANDOM FOREST WITH GRID SEARCH	82.405%

DECISION TREE WITH GRID SEARCH	**81.521%**
ADABOOST	**82.833%**

To predict income for individual above listed models in table have been used. It shows the overall accuracy for each model from that best model for prediction can be evaluate. In the table the accuracy for each model is shown, where the accuracy of each models is extremely close to each other. Out of all ADABOOST model outperform other model with the highest accuracy percentage which is 82.833% and the lowest is decision tree amongst all five of them.

V. **Conclusion**

In this paper we are focusing on identifying the salary class prediction for the better implementation of the salary models using the feature engineering and machine learning algorithm. Our business problem was to find out what type of adult makes more than $50K or less than $50K. We used visualization techniques to find that problem. From charts, we conclude that: adults who have mastered, or doctorate degrees makes more than 50K compared to those who have bachelor or associate degrees.

In the project, we have seen the highest accuracy was obtained from the **AdaBoost classifier Model**. The results were much better as we compared to work done in the past. The best accuracy we got on this dataset was **82.833%.** The accuracy of the models was close to each other but AdaBoost gave slightly better accuracy as compared to other two models which is used in this project.

For improvement on accuracy more work needs to be done to improve further. The future scope of this work requires to obtain an overall better set of results by using hybrid models with help of machine learning and deep learning together. By another way without reducing the accuracy applying other advanced pre-processing techniques.

Overall, our experience in this project as a team of three people was great because we learn how to work in a team and shared our knowledge with each other.

Thank you, Doctor Harry Wang to give us an opportunity to work at our pace and giving us freedom of choosing the dataset, and models.

REFERENCES

[1] https://archive.ics.uci.edu/ml/datasets/adult

[2] **Hands-on machine learning with Scikit-Learn & TensorFlow: concepts, tools, and techniques to build intelligent systems**
Géron - O'Reilly Media, Inc. – 2019

[3] **sklearn.ensemble.RandomForestClassifier¶**
https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[4] **AdaBoost Classifier in Python**
https://www.datacamp.com/community/tutorials/adaboost-classifier-python

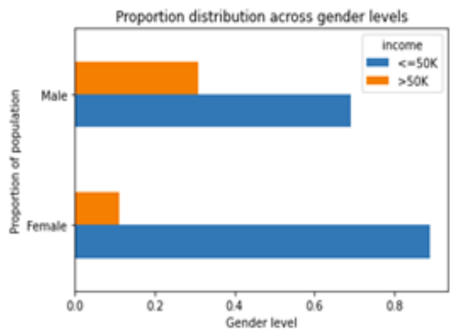
[5] **Your Machine Learning and Data Science Community**
https://www.kaggle.com/

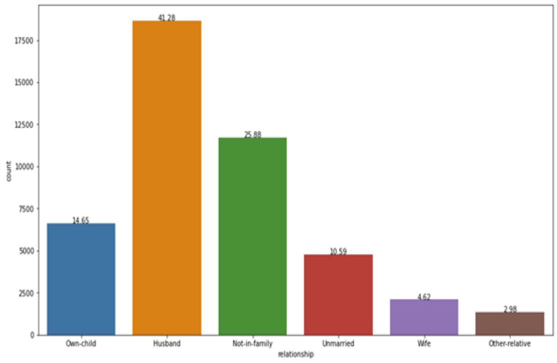
[6] **EDA of Adult Census Income Dataset**
Aamir
https://medium.com/data-warriors/eda-of-adult-census-income-dataset-cc9ac1a3d552

[7] **Statistical comparison of models using grid search¶**
https://scikit-learn.org/stable/auto_examples/model_selection/plot_grid_search_stats.html#sphx-glr-auto-examples-model-selection-plot-grid-search-stats-py

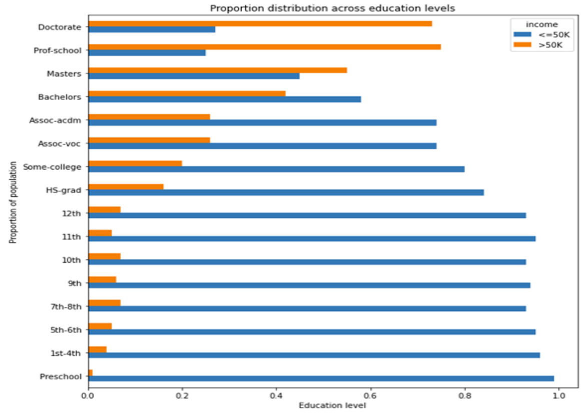
Appendix I

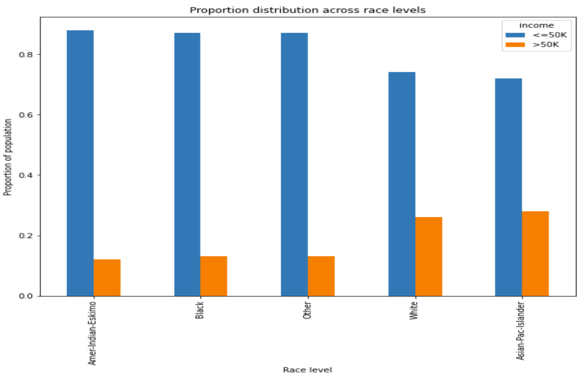
More Visualization Pictures:





Education vs Income


Proportion distribution across education levels
income
<=50K
>50K
Proportion of population
Doctorate
Prof-school
Masters
Bachelors
Assoc-acdm
Assoc-voc
Some-college
HS-grad
12th
11th
10th
9th
7th-8th
5th-6th
1st-4th
Preschool
Education level


Proportion distribution across race levels
income
<=50K
>50K
Proportion of population
Amer-Indian-Eskimo
Black
Other
White
Asian-Pac-Islander
Race level