

MACHINE LEARNING TECHNIQUES FOR CENSUS DATA

Vinay Chandragiri

Department of Computer Science

Indian Institute of Technology Guwahati, Guwahati, India.

ABSTRACT

This paper discusses about classification and regression models for the Census Data from the UCI Machine Learning Repository. The idea behind the paper is to ingest a dataset from the repository and gain some practical insights from it in order to develop suitable predictive models. This particular dataset has around 14 attributes and nearly 48000 instances. A binary classifier was developed from the available data that determines whether a person makes more than \$50k per year or not based on the census information. Some of the best data cleaning techniques were also explored for obtaining better understanding of the data. Scikit-Learn a python based Machine Learning library was used to obtain necessary results.

Keywords - Classification, Regression, Scikit-Learn, Census Data

I. INTRODUCTION

A Census is a procedure of acquiring people's information in a country and the way they are living.

These are the two major differences which make census a unique measure from other surveys, are that:

- 1) Everyone in the country is enumerated considering their economic conditions, source of income etc.
- 2) Also, everyone is enumerated at the same time which removes the bias

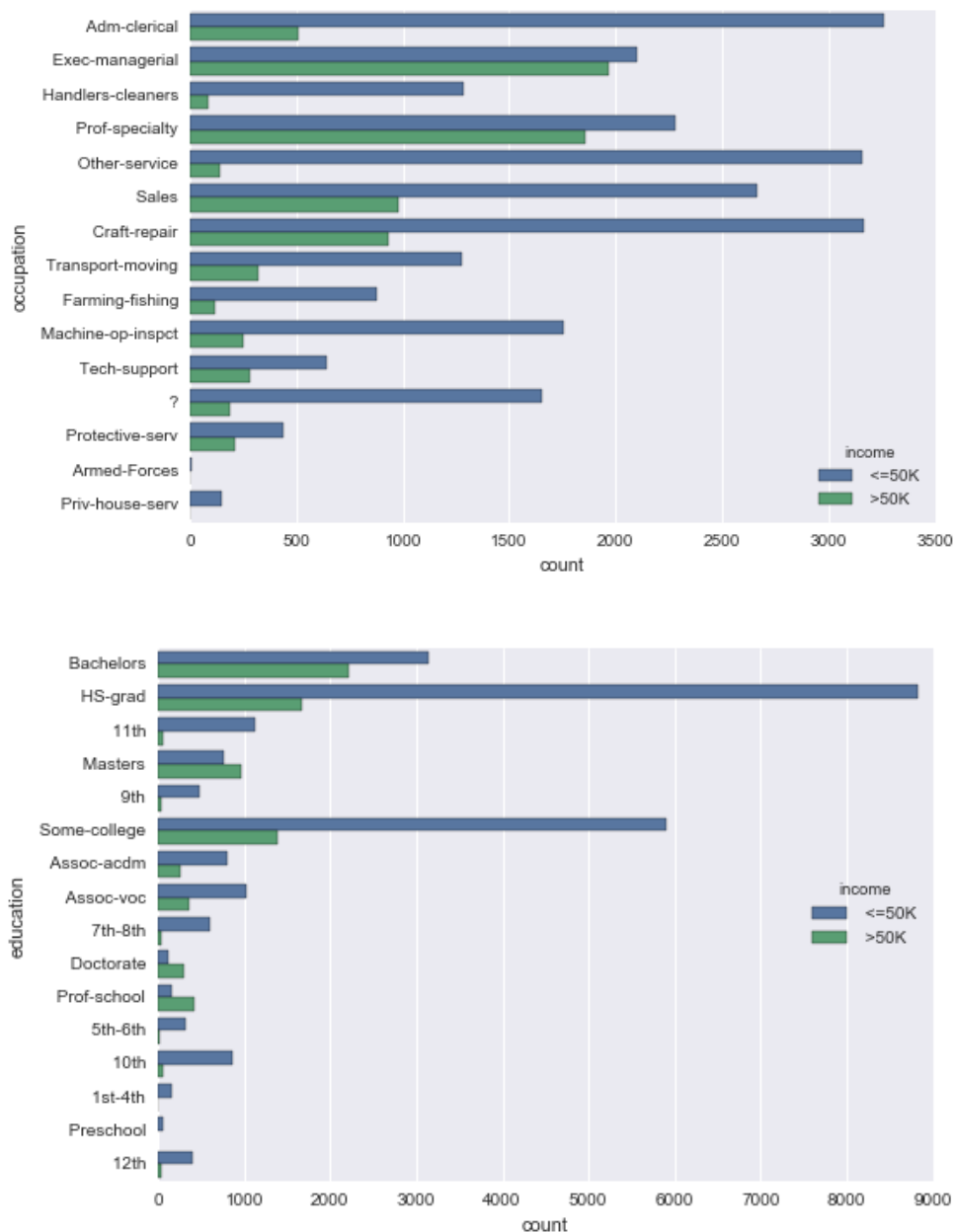
The census provides the government, businessmen, individuals and other interested people with information about the, social and economic characteristics, regarding housing and income, of the population. Without appropriate planning, economic development is not possible.

The questionnaire is developed in regard to Location, Demographic characteristics, Housing and Household facilities, Agriculture and fishing, Employment and Education, Publicity etc. Thus these are most of the attributes in our dataset.

II. DATASET

The Dataset used here is also called Adult dataset. The dataset contains three files with no file extension. They can be considered as text files if necessary. The adult.data file does not contain a header so manually, a header was constructed with all the attributes.

Here are the plots related to frequencies of "Occupation" and "Education" which are likely to predict the Income of an individual necessary for the task.



III. FEATURE EXTRACTION

Here we cannot use the categorical values for Machine Learning, so we need to encode into numerics which is one of the major part of Feature Extraction. Also, we need to impute missing values with data or else remove them for our good.

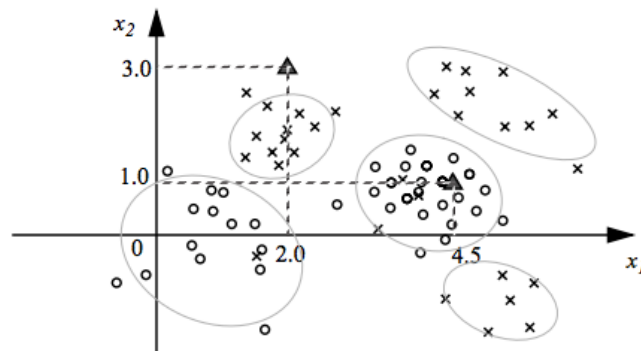
Scikit-Learn's Label encoder is used to encode categorical values as labels. As it can be used on a single vector over a time, we need to use it for all the columns in our dataset.

The second step left for us here is to remove unnecessary values such as "?" in the dataset. We'll have to ignore those rows. Imputer of sklearn.preprocessing was used to do the same but custom imputer can also be used for a multi-column approach. It is found that 'work-class', 'native-country' and 'occupation' were the only columns that require some sort of imputation.

IV. BUILDING THE MODEL

A pipeline was created in order to use the feature transformation and end up with some sort of classification. This pipeline helps in an developing an end to end system until using an estimator or classifier at the end by taking the data and transforming it. First the data gets passed through the encoder, then to the imputer and later to the classifier.

Logistic Regression decision boundary



An Illustration of Classification Task.

$$H(x) = 1 / 1 + e^{-x}$$

If $H(x) \geq 0.5$ or equivalently $x \geq 0$

Predict $Y = 1$

If $H(x) < 0.5$ or equivalently $x < 0$

Predict $Y=0$

V. RESULTS and CONCLUSION

A Logistic Regression classifier was build using which an F1 score of 0.8 was obtained. Also an SVM classifier was built using which an F1 score of 0.83 was obtained. However, other classifiers such as Naive Bayes and K-Nearest Neighbour can also be used. A simple gathering function is also developed that takes user input in order to predict the necessary.

This paper presents about an end to end analysis of a classification task of census data.

REFERENCES

- [1] Bache, K. & Lichman, M. (2013). UCI Machine Learning Repository [<http://archive.ics.uci.edu/ml>]. Irvine, CA: University of California, School of Information and Computer Science. Census Income Data Set, URL:<http://archive.ics.uci.edu/ml/datasets/Census+Income>
- [2] Hahsler, M. et al., Introduction to arules [Dash] A computational environment for mining association rules
- [3] Breiman, L. et al., Classification and regression trees, Chapman & Hall, 1984.