Prediction of Adult income

Master of Data Analytics

## Project report CISC 6930: Data Mining

Fall 2017

Submitted by:

Professor & Advisor: Yijun Zhao

Fordham University

**Index**

|  |  |
| --- | --- |
| Content | Page |
| 1. Overview |  |
| 1. Problem Statement |  |
| 1. Data analysis |  |
| 1. Data Pre-processing |  |
| * 1. Imputation |  |
| * 1. Data Encoding |  |
| 1. Training the model using data mining algorithms |  |
| * 1. . |  |
| 1. Model selection (Ensemble Method) |  |
| 1. Model application on testing set |  |
| 1. Conclusion |  |

1. ***Overview***

**Project Description**

This Data Mining project uses the real-world dataset extracted from 1994 census bureau database to predict annual incomes of adults, given a set of attributes like employment details, demographic information etc.

The income level is classified in two classes – less than 50,000 and greater than or equal to 50,000.

The project uses data various data mining algorithms and are written in python and R.

1. ***Problem Statement***

Using the given dataset with attributes, the aim is to build a predictive model that determines income level for adults. Income levels are classified in two classes below $50K and above $50K annually (given in the dataset).

1. ***Data Analysis***
2. Missing values:

After a preliminary exploration of the census data, it is found that both training and testing data sets contain missing value. For training and test data alike, all of the missing values were found to be confined in three categorical values: *native\_country, workclass and occupation*.

7.45% of training data instances, that is 2399 rows, contained missing values whereas 7.5% of test data instances, that is 1221 rows, contained missing values.



1. Unbalanced data:

Training data set was unbalanced with a negative skew, where 77% of the instances were classified as negative and only 23% were classified as positive.

1. Types of variables:

There are three types of variables found in the data set

1. Continuous:
2. Categorical:
3. Ordinal:
4. ***Data Pre-processing***
   1. Data cleaning and Imputation for missing values:

The training and testing datasets have data that is recorded as ‘?’. This needs to be managed so that it does not exert undue influence on the model. Two approaches are used for data cleaning and imputation

*A: Drop all the missing values*

In this approach, all missing values are dropped. The data is z-score normalized and ignored the unbalanced nature of the data in order to run preliminary tests and set a baseline for the classifier performance.

*B: Imputation*

Different techniques for imputation are used while still ignoring the imbalanced nature of data

1. Mode
2. Logistic Regression
3. Random Forest
4. K-nearest neighbors

*C: Select the imputation approach*

The most successful imputation approach is selected and data is balanced with the Bagging Classifier Method.

* 1. Data Encoding: