Aydan Alyanak
I526 Applied Machine Learning | Final Project
aalyanak@indiana.edu

Final Project

Empirical Question: In the programming assignment 1 and 2, we tried to predict whether a person has an income of less or more then 50K, in order to his/her demographics. While I was looking to the demographics, I came across with the fact that the income of Male attendees are more likely to be >50K then income of Female attendees.

My question is, is it because of women have less educated level or even though their educations are the same, does women get paid less then men?

For this final project, I compared the performance of Logistic Regression, Naïve Bayes, Decision Tree, Nearest Neighbor, SVM, Random Forest and Adaboost on the Adult data set from the UCI repository (http://archive.ics.uci.edu/ml/datasets/Adult).

The prediction task associated with this data set is to predict whether or not a person makes more than $50K a year using census data.

In addition to the prediction, I also looked at the possible reasons of my empirical question and tried to explain the outcomes.

1) Downloading the data via Python and Creating Weka Files

First, I used "requests" library of Python to connect to the website and download it, "pandas" to manipulate the data and "csv" library to read and write csv & arff files.

Training dataset = 'http://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data'

Test dataset = 'http://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.test'

The features:

AGE_GROUP (Discrete)	This is a categorized value, from a continuous age value to discrete. Values: {'< 25', '25-35', '35-45', '45-55', '55-70', '> 70'}
WORK_CLASS (Discrete)	This feature used as is. **NOTE: There were unknown (?) values, 70% of the known values are classified as "Private", therefore the unknown work_classes are also considered as "Private".**
FNLWGT (Continuous)	Final weight (This feature used as is)
EDUCATION (Discrete)	This feature used as is.
EDUCATION_NUM (Discrete)	Classes of the education (This feature used as is)
MARITAL_STATUS (Discrete)	This feature used as is.
OCCUPATION (Discrete)	This feature used as is. **NOTE: There were unknown (?) values, the unknown work_classes are considered as "Other-service".**
RELATIONSHIP (Discrete)	This feature used as is.
RACE (Discrete)	This feature used as is.
SEX (Discrete)	This feature used as is.
CAPITAL_GAIN (Continuous)	This feature used as is.
CAPITAL_LOSS (Continuous)	This feature used as is.
HOURS_PER_WEEK (Continuous)	This feature used as is.
CONTINENT (Discrete)	The native countries are remapped as continents. **NOTE: There were unknown (?) values, 90% of the known values are classified as "United-States", therefore the unknown work_classes are also considered as "United-States".** VALUES: {'Asia', 'Europe', 'North America', 'South America'}
INCOME (Class)	The class value.

2 arff files are created by Python, and ready to use. The rest of the analysis will be held in Weka.

2) Weka Analysis

a) Logistic Regression: Logistic regression is the appropriate regression (predictive) analysis to conduct when the dependent variable is binary. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. This is a useful approach to our dataset, and accuracies are expected to be high.

Correct classification rate = 85.60%
Incorrect classification rate = 14.39%

b) Naïve Bayes: The Naive Bayes Classifier technique is based on the so-called Bayesian theorem and is particularly suited when the dimensionality of the inputs is high [2]. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods. Since our dimensionality of inputs are not high, the method performed average.

Correct classification rate = 83.21%
Incorrect classification rate = 16.78%

c) Decision Tree: A decision tree is a graphical representation of possible solutions to a decision based on certain conditions [3]. It's called a decision tree because it starts with a single box (or root), which then branches off into a number of solutions, just like a tree. It's a very fitting approach to our dataset, as we can see in the result, it has the best performance.

Correct classification rate = 85.78%
Incorrect classification rate = 14.21%

d) K-Nearest Neighbors: k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification [4]. The k-NN algorithm is among the simplest of all machine learning algorithms. Both for classification and regression, it can be useful to assign weight to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones.

Correct classification rate = 83.21%
Incorrect classification rate = 16.78%

e) SVM: Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier [5]. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible.

Correct classification rate = 85.68%
Incorrect classification rate = 14.32%

f) Random Forest: Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classification or mean regression of the individual trees [6]. Random decision forests correct for decision trees' habit of overfitting to their training set.

Correct classification rate = 83.6%
Incorrect classification rate = 16.4%

g) Adaboost: The output of the other learning algorithms ('weak learners') is combined into a weighted sum that represents the final output of the boosted classifier [7]. AdaBoost is adaptive in the sense that subsequent weak learners are tweaked in favor of those instances misclassified by previous classifiers. AdaBoost is sensitive to noisy data and outliers. In some problems it can be less susceptible to the overfitting problem than other learning algorithms.

Correct classification rate = 82.4%
Incorrect classification rate = 17.6%

3) Results and Comparison

By aggregating all the results of 7 algorithms, the most likely fitting machine learning algorithm to this dataset appears to be Decision Trees, Logistic Regression and SVM.

	Correct Classification	Incorrect Classification	ROC Area	TP Rate	FP Rate	Precision
Decision Tree	**85.8%**	14.2%	0,867	**0,858**	0,329	0,851
Logistic Regression	**85.6%**	14.4%	0,908	**0,856**	0,317	0,850
SVM	**85.6%**	14.3%	0,766	**0,938**	0,407	0,882
Random Forest	**83.6%**	16.4%	0,879	**0,906**	0,391	0,882
Naïve Bayes	**83.2%**	16.8%	0,892	**0,832**	0,383	0,822
k-Nearest Neighbors	**83.2%**	16.7%	0,865	**0,832**	0,349	0,825
Adaboost	**82.4%**	17.6%	0,866	**0,978**	0,673	0,824

We can take the highest performance of correct classification (J48 - Decision Tree) and look at our empirical question, is it true that women are earning less than men, even if they have the same education level?

66% of the participants in the dataset are men, %34 are women.

Education level 9 (High School Graduates), 10 (College) and 13 (Bachelor) are the highest amount of participants.

- **Level 9:** %20 of men have the amount of >50K income, while 7% of the women have >50K.
- **Level 10:** %27 of men have the amount of >50K income, while 7% of the women have >50K.
- **Level 13:** %50 of men have the amount of >50K income, while 21% of the women have >50K.

There's a significant difference between women and men. The most high income of men have the occupations in "Craft-repair", "Exec-managerial", "Prof-specialty" and "Sales", we don't see high employments of women in these areas. Women are mostly distributed in all specialties, however the biggest employment of women is "Other-service" and get paid less than 50K, since we don't have a specific idea what an other-service is.

If we specifically look with the same education level of men & women in same job field, we'll also get interesting results. For example in Bachelor graduates:

- 28% of women in "Exec-managerial" position gets paid >50K, when 66% of men have that amount of income.
- 26% of women in "Sales" position gets paid >50K, when 55% of men have that amount of income.
- 24% of women in "Prof-specialty" position gets paid >50K, when 48% of men have that amount of income.

As a result, we can say "Women get paid less than men", and "Women don't get employed in the high income areas like men". That can be either employer's or women's choice.

REFERENCES

[1] http://www.statisticssolutions.com/what-is-logistic-regression/

[2] http://www.statsoft.com/Textbook/Naive-Bayes-Classifier

[3] http://study.com/academy/lesson/what-is-a-decision-tree-examples-advantages-role-in-management.html

[4] https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm

[5] https://en.wikipedia.org/wiki/Support_vector_machine

[6] https://en.wikipedia.org/wiki/Random_forest

[7] https://en.wikipedia.org/wiki/AdaBoost