# Predicting an Employee's wage – A comparison between Logistic Regression and Neural Networks

Falak Singhal
The University of Texas at Dallas, USA
fxs161530@utdallas.edu

Melvin James
The University of Texas at Dallas, USA
mxj162130@utdallas.edu

Vidya Sri Mani
The University of Texas at Dallas
vxm163230@utdallas.edu

## Abstract

*The aim of this project is to predict whether a given employee's salary will be lesser than or greater that 50K. A given person's salary depends on various factor's such as his location, educational qualification, age., The project uses two models, namely Logistic Regression and Neural Networks to classify the data (Adult data set) and use them to predict the salary of a new employee. The report provides a comparative study of the results. This project is completed using R.*

**Keywords:** Logistic Regression, Neural Networks, classification

## 1. Introduction

Employees, as well as job seekers are interested in knowing the salary compensation for a job, in some scenario. While job seekers are interested in knowing how much they can expect from a new job, existing employees may require a prediction of their compensation to understand the general wages. The wage of an employee is dependent on several factors, some more important that the other. A person with a certain job in one country may earn a wage which differs greatly if working in another country, for the same job.

Factors like location, Educational qualification, sex of the employee influence the pay-scale drastically. This project aims at predicting the pay of a new employee by learning from many other employees around the globe. It gives an idea about an employee's pay, whether it is less than 50K or exceeds 50K. Two classification models are used in this project, Logistic Regression and Neural Networks.

## 2. Overview


Fig 1: Project Overview

This project is carried out in three phases, Data collection, Model Fitting and prediction. Data collection uses the "adult" dataset. Following this phase, the models are trained to learn from the data provided. In the prediction phase, test data is tested against the two models.

## 3. Data Collection

This phase involves reading the data, and processing it to obtain better results.

### 3.1 Dataset description

This project uses the Adult dataset to train its models. "Adult" Dataset, also known as "Census Income" dataset. This data was extracted from the census bureau database found at http://www.census.gov/ftp/pub/DES/www/welcome.html.

| | |
|---|---|
| Data Set Characteristics | Multivariate |
| Number of Instances | 48842 |
| Attribute Characteristics | Integer, Categorical |
| Number of attributes | 14 |
| Number of instances with unknown values removed | 45222 |
| Number of instances unknown values removed | 3620 |

A total of fourteen attributes are mentioned in this dataset, seven of which are polynomials, one binomial and six continuous attributes. Seventy six percent of the records in the dataset have a class label of <50K.

Number of examples used for Training : 70% of total observations

Number of examples used for Testing : 30% of total observations

### 3.2 Reading the Data

Data is read from http://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data. The data set holds the following variables.

- age : The age of the individual
- type_employer : The type of employer the individual has. Whether they work for the government, military, private, etc.
- fnlwgt : The number of people the census takers believe that observation represents. This variable will be ignored.
- education : The highest level of education of that individual.
- education_num : Highest level of education in numerical form
- marital : Marital status of the individual
- occupation : The occupation of the individual
- relationship : Family relationship position, e.g., father, mother, etc.

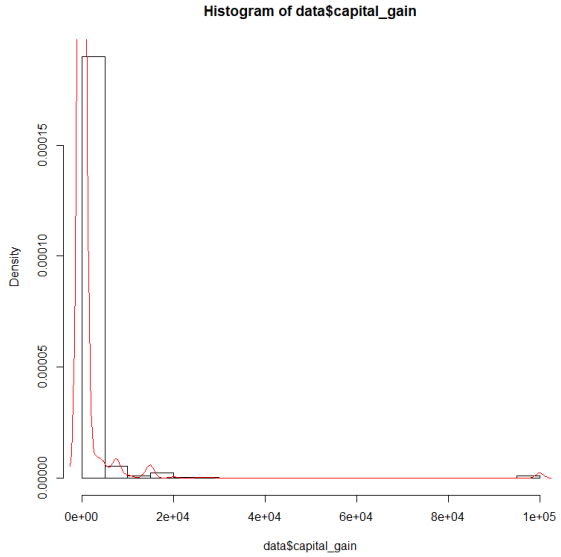
- race : descriptions of the individuals race. Black, White, etc.
- sex : Biological Sex
- capital_gain : Capital gains recorded
- capital_loss : Capital Losses recorded
- hr_per_week : Hours worked per week
- country : Country of origin for that person
- income : Boolean Variable. Whether that person makes more than $50,000 annually.

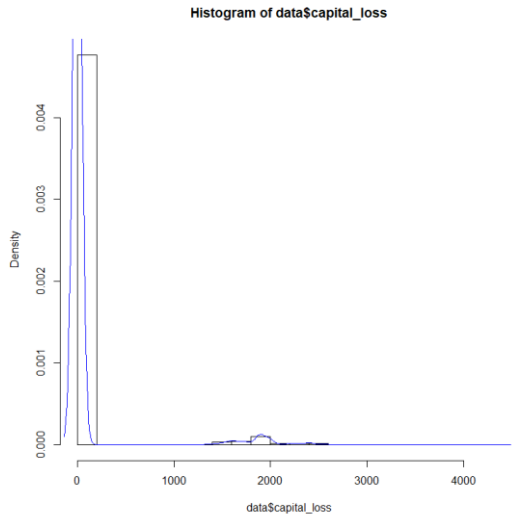
### 3.3 Cleaning and Processing the data.

Variable education_num clutters the analysis as it acts as a redundant variable, showing the numerical form of the education of a person.

Data stored as Text and Integers will be converted to vectors during import. Therefore, for each attribute, objects of type 'character' is defined.

The Frequency of values for some attributes may be too small, and hence their contribution might be ignored. To prevent this, smaller and related values are combined.




'Never-worked' and 'Without-pay' are both small and related groups and hence can be combined. Similarly, values in other attributes are combined.

For variable occupation, based on the industry, they are categorized as either blue or white collar.

Many entries in the data set have United states of America as their country. When compared to this, observations belonging to other countries are much smaller in number and hence their contribution would go unnoticed. To prevent this, countries are categorized geographically.

Values for a few other attributes are also combined. Observations are categorized as dropouts regardless of which grade they dropped out from. Besides the dropout category, the education variable combines other values in 'Masters', 'Doctorate', 'Associate' etc.,
Values are combines for variable marital and race as well.

Observations which don't have data for some values are omitted.

## 4. Fitting the Model

### 4.1. Cross Validation

Cross validation, popularly known as rotation estimation is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set.
A part of the complete dataset is removed for testing. Cross validation refers to the idea of splitting data as training and testing data.

The project uses 70% of the total number of observations for training purposes and the remaining 30% to test the data.

### 4.2. Logistic Regression

*4.2.1 Logistic Regression Theory*

Logistic Regression is a classification algorithm for fitting the regression curve, y=f(x), where y is a categorical variable. This algorithm predicts y, given a set of predictors, x.
y = [exp(b0 + b1x)] / [1 + exp(b0 + b1x)]

Logistic regression fits b0 and b1, the regression coefficients.

*4.2.2 Logistic Regression in R*

The glm() function is used to do generalized linear models.

Logistic Regression can be used, when we wish to predict a discrete variable, in this case, if the employee's wage is greater that 50K (represented as 1) or less than 50K (which we represent as 0), given a set of independent variables.

R provides support to fit a logistic regression model. The glm() function is used for the fitting process.

### 4.3. Neural Network

Neural Networks are used to identify non-linear patterns. Non-linear pattherns are those where one-to-one relationship between input and output is not present.

Neural networks work on the principle of identifying patterns between combinations of the input and the given output.

Prediction using Neural network in R requires the nnet Package.

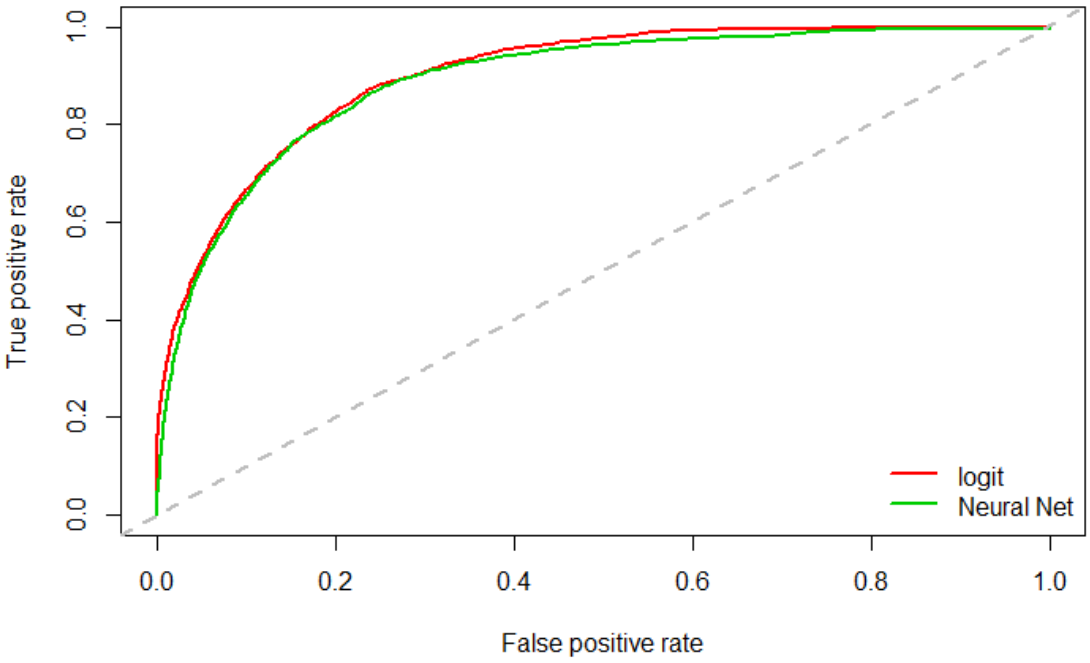
## 5. Prediction and Accuracy of Model on Test Data

### 5.1 True positives and false positives

A false positive error, commonly known as a 'false alarm' happens when a result that indicates a given condition has been fulfilled, when it has not. It is a type I error which takes reference from statistical hypothesis testing.
True positive refers to those scenarios where the criteria is marked as fulfilled when it fulfilled, i.e., the number of correct predictions.

### 5.2 ROC and AUC Analysis of fitted Models

The ROC Curve or Receiver Operating Characteristics curve is used estimate the performance of the Logistic Regression Model. It summarizes the model's performance by determining the tradeoff between true positives and false positive rate. The ROC curve summarizes the predictive capability




for all possible values of p > 0.5. The area under the curve (AUC), is referred to as index of accuracy(A) or concordance index. It can be used as a performance metric for ROC curve.
Higher the area under curve, better the prediction power of the model.

## 6. Accuracy of the two models

Logistic Regression ~ 85 %
Neural Network ~ 88 %

## References

Package 'nnet'. https://cran.r-project.org/web/packages/nnet/nnet.pdf

https://en.wikipedia.org/wiki/Cross-validation_(statistics)#k-fold_cross-validation

https://en.wikipedia.org/wiki/Machine_learning

https://en.wikipedia.org/wiki/False_positives_and_false_negatives#true_positive

https://www.r-bloggers.com/fitting-a-neural-network-in-r-neuralnet-package/

https://www.r-bloggers.com/evaluating-logistic-regression-models/

https://www.datacamp.com/community/tutorials/machine-learning-in-r#gs.zw49IYk

https://www.reddit.com/r/MachineLearning/

http://machinelearningmastery.com/machine-learning-in-r-step-by-step/

http://will-stanton.com/machine-learning-with-r-an-irresponsibly-fast-tutorial/