[Title: Predict Salary using Random](https://medium.com/@omairaasim/machine-learning-project-5-predict-salary-using-decision-tree-regression-cf1fe0cce169?source=your_stories_page---------------------------) Forest Algorithm



Name: Dipu Chandra Malo

ID:1712419042

Department of Electrical & Computer Engineering

North South University

Date:03/06/2020

Course Number : Cse445

Name: Machine learning

# Instructor’s Name: DR. MOHAMMAD ASHRAFUZZAMAN KHAN

# Abstract

The goal of this project is to predict salary of a person to some old data. Frist I entry a database then binary encoding the dataset. Then gives some vaule and random forest algorithm predict the salary. I also try to decision tree algorithm to predict the salary then try randomforest algorithm. This application can take the database for the salary system from the kaggle and makes a graph through this information from the database. It will check the salary fields then import a graph which helps to observe the database. And then it can predict a salary through the Random forest algorithm. It can also be applied in some other effective prediction also.

Table of Contents

[Abstract 2](#_Toc485290869)

[Table of Contents 3](#_Toc485290870)

[List of Figures 4](#_Toc485290871)

[List of Tables 5](#_Toc485290872)

[1 Introduction 6](#_Toc485290873)

[2 Background 8](#_Toc485290874)

[2.1 Subheading 1 7](#_Toc485290875)

[2.2 Subheading 2 7](#_Toc485290876)

[2.2.1 First Sub-subheading 7](#_Toc485290877)

[2.2.2 Second Sub-subheading 7](#_Toc485290878)

[3 Methodology 9](#_Toc485290879)

[4 Results and discussion 12](#_Toc485290880)

[5 Conclusions and recommendations 13](#_Toc485290881)

[6 Acknowledgements 14](#_Toc485290882)

[7 References 14](#_Toc485290883)

[Appendix A: Place the title of appendix here 16](#_Toc485290884)

# List of Figures

1 Corelation map ……………….1

2 Confusion Matrix of the classifier…………….. 1

# List of Tables

1.Data Pre-Procesing……………..1

2.Binary encoding output………………………1

# 1 Introduction

A prediction is an assumption about a future job salary prediction. A prediction is sometimes, though not always, is based upon knowledge or experience. Future events are not necessarily certain, thus confirmed exact data about the future is in many cases are impossible, a prediction may be useful to help in preparing plans about probable developments. In this paper salary of an employee of an organization is to be predicted on basis of previous salary database in use five feature is education.num ,occupation,age ,sex and marital.status . Here history of salary has been observed and then on basis of that salary of a person after encoding use of randomForest algorithm it can be calculated automatically.

In this paper the main aim is predicting salary and making a suitable user-friendly prediction of salary. From this prediction the salary of an employee can be observed according to a particular field according to their qualifications. It helps to see the growth of any field. It can produce a person’s salary by clustering and predict the salary through the graph. Using Ramdom Forest algorithm it makes prediction. Import matplotlib to get corelation map.This correlation map helps to predict the salary for any positions.

The application is aimed to develop to maintain a day-by-day monitoring to see the database(salary or experiences as well as designation, etc.). A confusion matrix of the classifer help us to predicted the salary.

A prediction is an assumption about a future event. A prediction is sometimes, though not always, is based upon knowledge or experience. Future events are not necessarily certain, thus confirmed exact data about the future is in many cases are impossible, a prediction may be useful to help in preparing plans about probable developments. In this paper salary of an employee of an organization is to be predicted onda. Here history of salary has been observed and then on basis of that salary of a person after a certain period of time it can be calculated automatically.

In this paper the main aim is predicting salary and making a suitable user-friendly graph. From this prediction the salary of an employee can be observed according to a particular field according to their qualifications. It helps to see the growth of any field. It can produce a person’s salary by clustering and predict the salary through the graph. Using linear regression and polynomial regression it makes a graph. This graph helps to predict the salary for any positions.

The application is aimed to develop to maintain a day-by-day monitoring to see the graphical medium of any field (salary or experiences as well as designation, etc.). A polynomial term: a quadratic (squared) or cubic (cubed) term turns a linear regression model into a curve. But because it is the data X that is squared or cubed, not the Beta coefficient, it still qualifies as a linear model. This makes it a nice and straightforward way to model curves without having to model complicated nonlinear models .One common pattern within machine learning is to use linear models trained on nonlinear functions of the data. This approach maintains the generally fast performance of linear methods while allowing them to fit a much wider range of data. That helps for the curving design.

It will help the employee as per following ways:

* Helping to see the growth at any field.
* With the help of machine learning it can easily produce a graph.
* Marketing easy to estimate the salary between x-y axis.
* User can give any point to get the salary through the program.
* Salary of the employees can be observed to give them a particular field according to their qualifications.

# 

# 2 Background

This prediction has been implemented through the following approaches:

* numpy:- NumPy is a Python package which stands for 'Numerical Python'. It is the core library for scientific computing, which contains a powerful n-dimensional array object, provide tools for integrating C, C++ etc. It is also useful in linear algebra, random number cap ability etc.

* matplotlib.pyplot:- matplotlib.pyplot is a collection of command style functions that make matplotlib work like MATLAB. Each pyplot function makes some change
* pandas:- pandas is a kind of software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license.

* read\_csv:- Python is a great language for doing data analysis, primarily because of the fantastic ecosystem of datacentric python packages. pandas is one of those packages and makes importing and analyzing data much easier.

* The random forest is a model made up of many decision trees. Rather than just simply averaging the prediction of trees (which we could call a “forest”), this model uses two key concepts that gives it the name random:
* Random sampling of training data points when building trees
* Random subsets of features considered when splitting nodes

First, a dataset in excel file to be made and then to be opened in Jupyter notebook from ANACONDA Navigator.From there we can read the datasetin ANACONDA Navigator. At Jupyter notebook first we taking three variables for importing purpose such as NumPy, matplotlib.pyplot and pandas. These functions are used are as follows:

* NumPy is used for antialiasing a dynamic array or large set.
* matplotlib.pyplot is used for making graph
* pandas are mainly used like database where we store the dataset from the excel file. Through the data set we are forming graph and predicting

Then importing the dataset on the pandas through the help of “<any variable>. read\_csv” and then showing that field that are importing from the excel file. Through the salary sets the plotting the points at the graph. After that I use random forest algorithm to predict the salary.

# 3 Methodology

Machine Learning (ML) is basically that field of computer science with the help of which computer systems can provide sense to data in much the same way as human beings do. In simple

words, Machine Learning is a type of artificial intelligence that extracts patterns out of raw data by using an algorithm or method. The key focus of Machine Learning is to allow computer systems to learn from experience without being explicitly programmed or human intervention. Random forest algorithm is a machine learning base algorithm .

Proposed Method for Salary Prediction:

Step 1: Salary data have been taken from dataset.

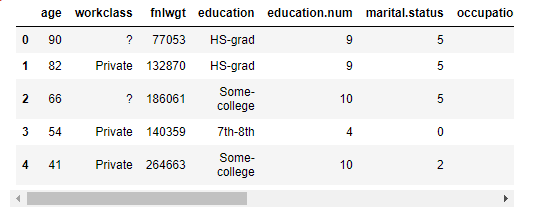
32561 entries, 0 to 32560

Data columns (total 15 columns):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Number | Column | Non-Null Count | Datatype |  |
| 0 | Age | 32561 non-null | Int64 |  |
| 1 | Workclass | 32561 non-null | Object |  |
| 2 | Fnlwgt | 32561 non-null | Int64 |  |
| 3 | Education | 32561 non-null | Object |  |
| 4 | Education.num | 32561 non-null | Int64 |  |
| 5 | Marital.status | 32561 non-null | Object |  |
| 6 | Occupation | 32561 non-null | Object |  |
| 7 | Relationship | 32561 non-null | Object |  |
| 8 | Race | 32561 non-null | Object |  |
| 9 | Sex | 32561 non-null | Object |  |
| 10 | Capital.gain | 32561 non-null | Int64 |  |
| 11 | Capital.loss | 32561 non-null | Int64 |  |
| 12 | Hours.per.week | 32561 non-null | Int64 |  |
| 13 | Native.country | 32561 non-null | Object |  |
| 14 | income | 32561 non-null | Object |  |

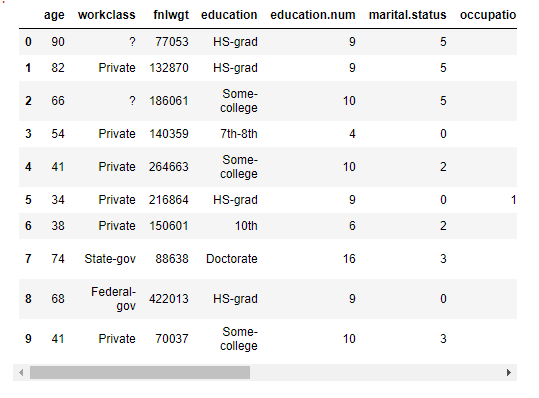
datatypes: int64(6), object(9)

Step 2: Data Pre-processing in five feature.



**Data Pre-Procesing**

Step 3: Binary encode the output of data pre-processing.



**Binary encoding output**

Step 4:Corelation map gives us idea.

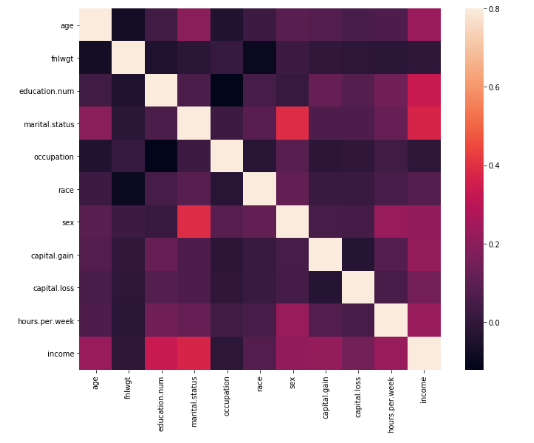
Step 5: Confusion matrix of the classifier.

Step 6:Setting up random forest algorithm.

Step 7: After that we using random forest algorithm.

Step 8: Give the 5 value of education.num ,occupation,age ,sex and marital.status in number.

Step 9: Predic the salary is smaller then 50 k or higher than 50k.



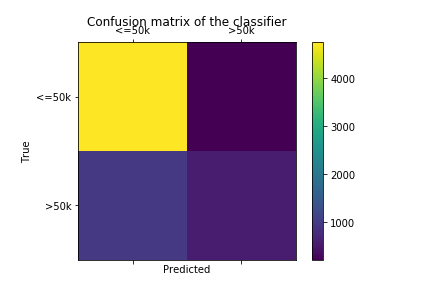
**Figure.1:Corelation map**

**4 Results and discussion**

From this application can be observed as the graphical representation and can also predict the automatically salary. Also survey the salary venue. Surveys of salary are therefore differentiated on basis of their data source into those that –

* Get data from companies, or
* Collect data from employees. Survey operators assigned for salary strive to get the most significant input data in every possible way. There is no way to decide that which approach is correct. The first possibility may assure large companies, where as the second choice is mainly for comparative smaller companies.
* When use decision tree algorithm *Predicting the result and giving the accuracy*  0.7997850452940273 and when use random forest algorithm *Predicting the result and giving the accuracy*  0.8182097343773991.
* We see the different is too much low.but random forest algorithm give petty much more

accuracy prediction of salary .



# Figure.2:Confusion Matrix of the classifier

# 5 Conclusions and recommendations

* Choosing the salary from predict use the random forest algorithm.i used correlation map. It will depend on the way of choosing the way to present the data, as well as its own preferences. Now a days spreadsheet programs like Excel are very flexible to create map of different types; with a few number of clicks one can see the represented data as a bar map,or a circle graph. This prediction is correct upto a certain percentage. More accurac obtained by implementing random forest algorithm. Decision tree algorithm petty much low accuracy .From there the best prediction can be chosen. It can be improved in following way:
* It can give more advance software for tallying salary medium.
* It will host the platform on online servers.
* It can do as large database also and curve also bigger than above example.

This predictor method can be used for predicting population of a country as well as forecasting a daily issue.

# 6 Acknowledgements

This work was conducted, under by DR. MOHAMMAD ASHRAFUZZAMAN KHAN. The authors are also particularly grateful to DR. MOHAMMAD ASHRAFUZZAMAN KHAN for his class lecture and slide helps me to make it.I collect data form kaggle .

# 7 References

1. Andreas Mullar, “Introduction to Machine Learning using Python: A guide for data Scientist,” in O’Reilly Publisher, India.

2. S. Marsland, Machine learning: an algorithmic perspective. CRC press, 2015.

3. A. L. Buczak and E. Guven, “A survey of data mining and machine learning methods for cyber security intrusion detection,” IEEE Communications Surveys & Tutorials, vol. 18, no. 2, pp. 1153–1176, Oct., 2015.

4. Tzanis, George, et al. "Modern Applications of Machine Learning." Proceedings of the 1st Annual SEERC Doctoral Student Conference–DSC. 2006.

5. Horvitz, Eric. "Machine learning, reasoning, and intelligence in daily life: Directions and challenges." Proceedings of. Vol. 360. 2006.

6. Mitchell, Tom Michael. The discipline of machine learning. Carnegie Mellon University, School of Computer Science, Machine Learning Department, 2006.

7. Arum, R. (1998). The effects of resources on vocational student educational outcomes: Invested dollars or diverted dreams? Sociology of Education, 71, 130-151. 8. Lewis, C. D., 1982. Industrial and Business Forecasting Methods, London, Butterworths.

# Appendix A: Place the title of appendix here

**A.1 Motivation of the Random Forests approach**

While decision trees are very effective at modeling complex data relationships, they have two notable drawbacks. First, individual decision trees may have a tendency to be overfit—in other words, they conform too closely to the particular samples represented in the training dataset. While this makes their fit (in terms of R2 or similar metrics) very good, it renders them less accurate at classifying new samples. In addition, decision trees on their own can only be used in classification problems; they cannot be applied to clustering problems like the one addressed in this study, where the class identities of training samples are unknown. RF is an extension of the decision tree framework that addresses both of these limitations. Put simply, RF produces a collection of numerous decision trees (i.e., a forest), each based on a different random permutation of the training data. Results from all trees in the collection are averaged to make predictions, rather than allowing any one tree to dictate the analysis. The randomness in the procedure reduces overfitting, leading to more robust predictions. In addition, the RF algorithm provides a powerful measure of relatedness or proximity among samples. As discussed below, this last feature allows RF to be used in clustering problems. This study employs the RF algorithm first introduced by Breiman (2001), which is the most widely used variant today. In his foundational introduction of the method, Breiman showed that it has accuracy at least as good as other ensemble decision tree algorithms, is robust to outliers and noise, and is computationally efficient compared to alternatives. The basic algorithm is implemented using the randomForest R package (Liaw et al. 2015), and the RF clustering approach is based on recommendations of Shi and Horvath (2006).

**A.2 The Random Forests algorithm**

Consider a dataset of N training samples consisting of the N × 1 response vector of categorical class labels y, along with K predictor variables contained in the N × K matrix X. To create a model for predicting y from X, the RF algorithm first creates M trees (where M is typically on the order of 103 –104 ), with randomness injected among them at two different levels. First, each tree is constructed from a random draw of N observations selected with replacement from the training dataset. That is, for each tree a vector φ of N indices is first drawn randomly (with replacement) from the integers [1, N]. Then, the responses and predictors to be fit by the tree, ! and ! (respectively), are defined as: Final BSLs Report – Appendix A 22 January 2016 A-3 ! = !!!, !!!, … !!! (2) ! = !!!,! ⋯ !!!,! ⋮ ⋱ ⋮ !!!,! ⋯ !!!,! (3 This scheme allows individual training samples to contribute more or less to model fit in each of the M random decision trees. This property prevents unusual samples or spurious relationships among a few particular samples from dominating the final outcome of the model. The second random element in the Breiman RF algorithm enters at the node level. Unlike the single-tree analysis described above in Section A1.2, individual trees in an RF analysis are not formed by exhaustively searching for the best split at every node. Instead, for each split the algorithm considers !! predictor variables randomly selected from the K available predictors in X (typically, !! ~ !), and chooses the best possible split from among these K’ predictors. The result of this additional randomization is that different variables are considered in each tree and each node, which allows for a wider range of influences from each variable on the final model. For example, early splits on the most important predictors might dominate trees built with the full X, consistently overwhelming substantial but comparatively subtle effects of one or more other predictors. By randomly omitting the dominant predictors from some splits, these effects will be evident in at least some of the trees in the forest.