

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/339055809>

Salary Prediction Using Regression Techniques

Article in SSRN Electronic Journal · January 2020

DOI: 10.2139/ssrn.3526707

CITATIONS

9

READS

15,317

3 authors, including:



Rupashri Barik

JIS COLLEGE OF ENGINEERING

10 PUBLICATIONS 12 CITATIONS

[SEE PROFILE](#)



Ayush Mukherjee

Technische Universität Ilmenau

1 PUBLICATION 9 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



Digital Image Processing [View project](#)



Software Engineering [View project](#)

SALARY PREDICTION USING REGRESSION TECHNIQUES

MANUSCRIPT TRACK: MACHINE INTELLIGENCE

Sayan Das(JIS College of Engineering, Kalyani, Nadia), Rupashri Barik*(JIS College of Engineering, Kalyani, Nadia), Ayush Mukherjee(JIS College of Engineering, Kalyani, Nadia).

Abstract:

The goal of this paper is to predict salary of a person after a certain year. The graphical representation of predicting salary is a process that aims for developing computerized system to maintain all the daily work of salary growth graph in any field and can predict salary after a certain time period. This application can take the database for the salary system from the organisation 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 graphical representation. And then it can predict a certain time period salary through the prediction algorithm. It can also be applied in some other effective prediction also.

Keywords:

Machine Learning, Linear Regression, Polynomial Regression

*Email: rupashri.barik@jiscollge.ac.in

Phone: 9830197914

I. INTRODUCTION

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 on basis of previous salary growth rate. 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.

The graphs through the Linear and polynomial graphs are displayed to detect the salary and position levels.

II. METHODOLOGY

Machine Learning (ML) [1] 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, ML [2] [3] is a type of artificial intelligence that extracts patterns out of raw data by using an algorithm or method. The key focus of ML is to allow computer systems to learn from experience without being explicitly programmed or human intervention.

Linear Regression [4] is an algorithm of machine learning based on supervised learning scheme. Linear regression [6] carries out a task that may predict the value of a dependent variable (y) on basis of an independent variable (x) that is given. Therefore, this kind of regression technique looks for a linear type of relationship between input x and output y .

Polynomial Regression [5] [7] is a form of linear regression in which the relationship between the independent variable x and dependent variable y is modelled as an n th degree polynomial. Polynomial regression [8] fits a nonlinear relationship between the value of x and the corresponding conditional mean of y , denoted $E(y|x)$.

This application provides a Salary graph representation that is mainly done by polynomial regression statistics, polynomial regression [8] is a form of regression analysis which represents the relationship between the independent variable x and the dependent variable y and that is modelled as the n th degree of polynomial in x . Polynomial regression is suitable for a nonlinear type of relationship between the value of x and the correlating conditional mean of y , represented as $E(y|x)$. Although polynomial regression fits a nonlinear model to the data, as a statistical estimation problem it is linear, in the sense that the regression function $E(y|x)$ is linear in the unknown parameters that are estimated from the data. For this reason, polynomial regression is assumed to be a special case of multiple linear regressions.

Curve fitting is a method of building a curve, or to represent a mathematical function that is optimally suitable to a series of data points, and possibly it is subject to constraints. Curve fitting can involve either interpolation, where an exact relevant to the data is needed, or smoothing, in which a "smooth" function is created that is approximately suitable the data. A related topic is regression analysis, which focuses more on questions of statistical inference such as how much uncertainty is present in a curve that is fit to data observed with random errors. Fitted curves can be

used as an aid for data visualization, to infer values of a function where no data are available, and to summarize the relationships among two or more variables. Extrapolation refers to the use of a fitted curve beyond the range of the observed data, and is subject to a degree of uncertainty since it may reflect the method used to construct the curve as much as it reflects the observed data.

Following is the descriptions of the method that the work has been done.

Proposed Method for Salary Prediction:

Step 1: Salary data have been taken from dataset.

Step 2: Then the points corresponding to the salary data of an individual person have been plotted in the graph. The data are initialized in pandas (ascending, descending, mixed-up). Taking the dataset from each pandas field and from the pandas dataset we plotted the points on the graph as per number wise or input wise that came real dataset.

Step 3: After that we using linear regression for draw lines between the points.

Step 4: If the points are not in linear way then we use polynomial regression for curving purpose. Through the clustering points we can make a smooth and curve path.

Step 5: After then through the linear/polynomial graph through the x - y axis we can predict salary.

Step 6: Also, we predict a person on future salary position as per the graph goes. Only take a particular person position, then the prediction answer be executed through the help of the graph.

III. EXPERIMENTATION

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

- to a figure: e.g., creates a figure, creates a plotting area in a figure, plots some lines in a plotting area, decorates the plot with labels, etc.
- 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 data-centric python packages. pandas is one of those packages and makes importing and analyzing data much easier.
- Linear Regression :- Linear Regression will be used to perform linear and polynomial regression and make predictions accordingly. Now, you have two arrays: the input x and output y. You should call .reshape() on x because this array is required to be two-dimensional, or to be more precise, to have one column and as many rows as necessary,
- Polynomial features: - Generate a new feature matrix consisting of all polynomial combinations of the features with degree less than or equal to the specified degree. For example, if an input sample is two dimensional and of the form [a, b], the degree-2 polynomial features.
- Predict:-predict the values where find at the time period that how much there in that graph.

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 dataset in 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 then it is being tried to input linear regression function makes only straight line, if all points are linear then these function is used and directly goes making graph and predicting. But if the points are not linear so we go for next step. If the points are not linear then we use polynomial function for curving purpose that help to observing user that how the growth are moving on a path. After this the diagram on the graph is for demonstration. Then it is predicted to a salary through the graph using both the x-y-axis. "<variable>.predict(poly_reg.fit_transform(v))". From here it gets a prediction value through the polynomial/linear regression.

IV. RESULTS AND DISCUSSION

From this application can be observed as the graphical representation and can also predict the any point from position and automatically calculates 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.

Output of accessing data set is as follows:

Out[2]:

	Position	Level	Salary
0	Business Analyst	1	45000
1	Junior Consultant	2	50000
2	Senior Consultant	3	60000
3	Manager	4	80000
4	Country Manager	5	110000
5	Region Manager	6	150000
6	Partner	7	200000
7	Senior Partner	8	300000
8	C-level	9	500000
9	CEO	10	1000000

Figure 1 Sample Dataset

First we linear function. “LinearRegression()” from these function it makes straight line through the points.

Output:

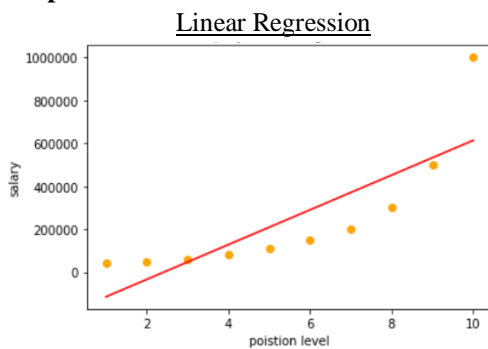


Figure 2: Results of linear regression for the salary dataset

Now, plotting the point where x-axis represents the position and y-axis presents the salary from database Position_Salaries.

Output:

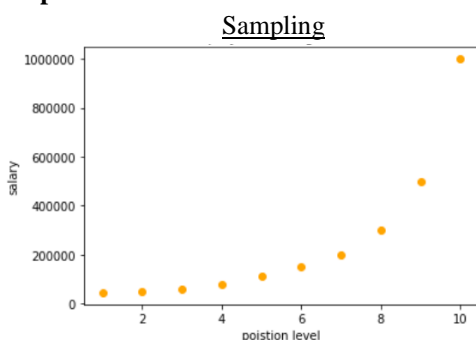


Figure 3: Sampling the data from the salary dataset

In Figure 3 we can see all the points are not using. For that reason now we are implementing “PolynomialFeatures” for curviness. “degree = 6” here refer the smoothness of curve .

Same like above here, “lg2.predict(poly_reg.fit_transform(x))” function is used for curving than straight line.

Output:

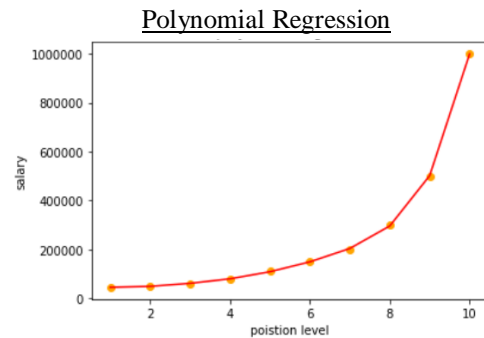


Figure 4: Results of polynomial regression for the salary dataset

V. CONCLUSIONS

Choosing the salary from x-y graph as it is to use to represent a specific data set that takes some trial and error. Often there may be more than one appropriate type of graph to be used. 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 graphs of different types; with a few number of clicks one can see the represented data as a bar graph as well as by a line graph, or a circle graph. This prediction is correct upto a certain percentage. More accuracy can be obtained by implementing k-nearest regression. 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.

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.

Author Biographical Statements

<p style="text-align: center;">Sayan Das</p> <p>Sayan Das is a student of BCA, 3rd year of JIS College of Engineering, Kalyani, Nadia. His main area of interest is Machine Learning.</p> 
<p style="text-align: center;">Rupashri Barik</p> <p>Rupashri Barik is working as Assistant Professor at Dept. Of Information Technology, JIS College of Engineering, Kalyani, Nadia. She has teaching experience of more than 12 years. Her main areas of interests are Digital Image Processing, Machine Learning.</p> 
<p style="text-align: center;">Ayush Mukherjee</p> <p>Ayush Mukherjee is a student of BCA, 3rd year of JIS College of Engineering, Kalyani, Nadia. His main area of interest is Machine Learning.</p> 