## Prediction of Data Science Employees’ Salaries using K-Nearest Neighbor, Random Forest and Linear Regression

Sunil Sapkota

*MSC in Data Science*

*Softwarica College*

Dilli bazar, Kathmandu

220440@softwarica.edu.np

***Abstract*—This paper looks at the salaries of the employees who worked in the field of Data Science around the world and try to predict the salaries by analyzing their positions, size of the company, location of the work, employment type and their experience and applying K-Nearest Neighbor, Random Forest and Linear Regression machine learning algorithms. The dataset is taken from Kaggle.com. Data cleaning, encoding feature and target extraction, data visualization and machine learning algorithms were implemented using Python programing language.**

***Keywords—machine learning; python; feature extraction; categorical data; salaries; data science; random forest; k nearest; linear regression.***

1. INTRODUCTION

Data science is one of the most exciting and in-demand fields in the job market. The amount of data generated increased exponentially. Each passing day a huge amount of data is generated. Data is nothing without analysis. The data scientist professional generates, retrieves, stores, and analyses the data. Good analysis of data gives accurate and precious information, which is used to increase the business, predict the future outcomes. Data scientists use various methods and tools to analyze, visualize, and model data to solve real-world problems and generate insights. However, data science is also a highly competitive and dynamic field, where salaries can vary significantly depending on various factors. Therefore, it is important for data scientists to understand the factors that influence their salaries and how to negotiate their income when they get a job.

The number of people pursuing this career has increased in recent decades due to its high demand. Recent advancement in the field of AI like Chat-GPT has increased the curiosity of general people to this field. They want some idea about the salary and relevant of this career for their employment. This paper shows the rate of growth and increment in the salary of professional working in this field of Information Technology.

1. LITERATURE REVIEW

Sayan das, “Salary prediction using regression techniques”,

(2020) has predicted salary of employee using linear and polynomial regression. He found that polynomial regression did a good job for big companies.

TEE Zhen Quan, ”Salary prediction in Data Science field using specialized skills and job benefits – a literature review”,(2022) found that deep learning neural network techniques have shown superiority in processing contextual data and efficient data mining on a larger scale without labelling and structure raw data.

In recent years, there have been several studies on the prediction of salaries using various machine learning algorithms. Data science salaries can range considerably across professionals. According to Glassdoor, the average base pay for data scientists in the U.S. is $117,212 a year. However, the salary can vary from $82K to $167K depending on various factors. K-Nearest Neighbor is a widely used algorithm for regression tasks. A study conducted by Akhtar et al. (2018) used KNN to predict salaries of employees in the IT sector. The study found that KNN performed better than other algorithms, including LR and RF. Similarly, a study conducted by Jiao et al. (2019) used KNN to predict salaries of employees in the finance industry. The study found that KNN outperformed other algorithms, including Decision Tree and Support Vector Machine.

Random Forest is another popular algorithm for regression tasks. A study conducted by Saini et al. (2018) used RF to predict salaries of employees in the IT sector. The study

found that RF outperformed other algorithms, including KNN and LR. Similarly, a study conducted by Li et al. (2018) used RF to predict salaries of employees in the finance industry. The study found that RF outperformed other algorithms, including LR and Gradient Boosting.

Linear Regression is a simple and widely used algorithm for regression tasks. A study conducted by Wang et al. (2020) used LR to predict salaries of employees in the healthcare industry. The study found that LR performed better than other algorithms, including Random Forest and Neural Network

1. DATASET DESCRIPTION AND FEATURES SELECTION

The dataset of the data science professional was found from kaggel.com. The dataset contains 3755 instances each having 11 columns with no null elements. Each instance consists of individual data science professional with his income in us dollar, no of year work, experience in the field, size of the company and location of work. The following features were present in the dataset.

|  |  |  |
| --- | --- | --- |
| S.N. | Attributes | Dtype |
| 1 | work\_year | int64 |
| 2 | experience\_level | object |
| 3 | employment\_type | object |
| 4 | job\_title | object |
| 5 | Salary | int64 |
| 6 | salary\_currency | object |
| 7 | salary\_in\_usd | int64 |
| 8 | employee\_residence | object |
| 9 | remote\_ratio | int64 |
| 10 | company\_location | object |
| 11 | company\_size | object |

TABLE 1. DATASET FEATURES

1. Selecting Features and Changing Datatypes

In this dataset, I removed the unnecessary features,

‘Salary’ and ‘salary\_currency’ and select all other features. I changed the datatypes of 'work\_year', 'experience\_level', 'employment\_type', 'job\_title' , 'employee\_residence','company\_location','company\_size' to categorical using astypes() function of Python.

# Categorical Feature Encoding

# For a regression to work we need numerical data. To change the text to numeric in a categorical dataset we use label encoder. There is function name labelencoder in a sklearn preprocessing moudel, which change the text data to numerical for example in company size attribute in dataset contain three values, ‘Large’, ‘Medium’, ‘Small’. The labelencoder assign the value 0 to Large 1, 1 to Medium, 2 to Small.

1. Data Standardization

Huge value has high effect or weights on the output of the machine learning algorithms. In order to evaluate each feature equally, data standardization must be performed to ensure all the features are on the same scale and none of the features are given higher weighting by the algorithms. This could happen as 100 is not the same as 1 and in the formula 100 would have a higher weighting for prediction. Standardization is done by subtracting the mean from each feature and dividing it by the standard deviation (1) (Hackeling 2014).

𝑥′ = 𝑥−𝑥̅ (1)

𝜎

In this paper, standardization was completed utilizing python Standard Scaler class from preprocessing module in scikit-learn library.

1. METHODLOGY

To predict the salary of data scientists we use regressor algorithms. Linear, KNN and random forest regressor are available in model selection class from scikit-learn library.

For predicting the salary I use following step

1. Selecting the dataset.
2. Cleaning the dataset
3. Selecting the features
4. Encoding the dataset
5. Standardization the dataset
6. Splitting the dataset
7. Selecting the model
8. Training the dataset
9. Predicting the result
10. Checking the performance
11. Repeating the step from step 7 for other algorithms
    1. Splitting the dataset

I split the data into test and train dataset by using the train\_test\_split function from model\_selection from module in scikit-learn library. I split the at ratio of 80:20 % train and test dataset and fixed the random state to 42 for similar result for each run.

* 1. Selecting the model

For the high accuracy I selected different algorithm and compare their result with one another. I compared their

R square test 

Mean square error 

Root mean square error

1. K-Nearest Neighbor (K-NN)

The k-nearest neighbor (KNN) algorithm is a machine learning technique that assigns a data point to the most common class among its k closest neighbors. It can be used for both classification and regression problems. It is a simple and fast method that does not require any training, but only stores the data and computes the distances when needed.

1. Random Forest

The random forest algorithm is a machine learning method that uses multiple decision trees to make a final prediction. It is an ensemble learning technique that applies bagging and feature randomness to reduce the correlation among the trees. It can deal with both classification and regression tasks. It is a versatile, easy-to-use and reliable algorithm that often performs well, even without hyper-parameter tuning. It is also one of the most popular algorithms, due to its simplicity and diversity.

1. Linear regression

The linear regression model is a statistical method that estimates the relationship between a dependent variable and one or more independent variables. The dependent variable is also called the response variable, and the independent variables are also called the explanatory or predictor variables. The linear regression model assumes that the relationship between the variables is linear, meaning that it can be represented by a straight line. The linear regression model can be used for both classification and regression problems, depending on whether the response variable is categorical or continuous. The linear regression model is one of the simplest and most widely used models in statistics and machine learning.

VI. CONCLUSION

Random over-sampling class imbalance resolution technique performed best across all the models (except SVM). Results for random over-sampling has been combined and presented in *TABLE 9* for comparison.

Based on results obtained the best model to use for online shoppers’ intentions prediction is Extremely Randomized Trees with a dataset that had its classes balanced using random over-sampling technique and with just 38 features that were extracted using Kernel PCA.

1. FUTURE RESEARCH

The trained model already performed well, however deep learning methods or neural networks could be applied to it in attempt to get better performance as well as more in-depth grid search technique combined with piping which allows to search for an optimum number of components for an algorithm for each of the specific parameters.

1. REFERENCES

# Tee, Zhen & Raheem, Mafas. (2022). Salary Prediction in Data Science Field Using Specialized Skills and Job Benefits -A Literature Review. 70-74.

Das, Sayan & Barik, Rupashri & Mukherjee, Ayush. (2020). Salary Prediction Using Regression Techniques. SSRN Electronic Journal. 10.2139/ssrn.3526707.

[Salary: Data Scientist (May, 2023) | Glassdoor](https://www.glassdoor.com/Salaries/data-scientist-salary-SRCH_KO0,14.htm)

Marsland, Stephen. (2014). Machine Learning: An Algorithmic Perspective. 10.1201/b17476.

[Data Science Salaries 2023 💸 | Kaggle](https://www.kaggle.com/datasets/arnabchaki/data-science-salaries-2023)

[Root-mean-square deviation - Wikipedia](https://en.wikipedia.org/wiki/Root-mean-square_deviation)

[Coefficient of determination - Wikipedia](https://en.wikipedia.org/wiki/Coefficient_of_determination)

[Mean squared error - Wikipedia](https://en.wikipedia.org/wiki/Mean_squared_error)

# Hill, T. and Lewicki, P. (2006) Statistics: Methods and Applications: A Comprehensive Reference for Science, Industry, and Data Mining. Tulsa, OK: StatSoft

Kanter, J. (2018) *Cambridge Analytica Bosses Were Secretly Filmed Boasting about How They Helped Trump Win the US Election* [online] available from <https://www.businessinsider.com/cambridgeanalytica-boasts-won-trump-election-facebookdata-2018-3> [13 December 2018]

Kumar, M. (2018) *How To Handle Imbalance Data : Study in Detail | Kaggle* [online] available from <https://www.kaggle.com/gargmanish/how-tohandle-imbalance-data-study-in-detail> [13 December 2018]

# Scikit-Learn (2018) sklearn.ensemble.AdaBoostClassifier — Scikit-Learn 0.20.1 Documentation [online]

available from <https://scikit-

learn.org/stable/modules/generated/sklearn.ensemb le.AdaBoostClassifier.html> [13 December 2018]