## **GRAPHIC ERA DEEMED TO BE UNIVERSITY**

## **DEHRADUN**

****

**MINI PROJECT REPORT**

**On**

**Machine Learning Based Application**

**VINAY KUMAR**

**2015401**

**CST**

**30**

**DATED - 14/12/2021.**

**Abstract –**

Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.

Machine learning is an important component of the growing field of data science. Through the use of statistical methods, algorithms are trained to make classifications or predictions, uncovering key insights within data mining projects. These insights subsequently drive decision making within applications and businesses, ideally impacting key growth metrics. As big data continues to expand and grow, the market demand for data scientists will increase, requiring them to assist in the identification of the most relevant business questions and subsequently the data to answer them.

So, here I have used data science to predict the salaries of developers with different experience level, college degree, country and predicted the average salary for the user. To get the least error, I have used many models to predict the proper value with least number of errors.

This project will help us in finding what salary we should expect from other companies, so rather than thinking ourselves, we can take help from this application.

**Introduction –**

Machine Learning, as the name says, is all about machines learning automatically without being explicitly programmed or learning without any direct human intervention. This machine learning process starts with feeding them good quality data and then training the machines by building various machine learning models using the data and different algorithms. The choice of algorithms depends on what type of data we have and what kind of task we are trying to automate.

Now days, Major reason an employee switches the company is the salary of the employee. Employees keep switching the company to get the expected salary. And it leads to loss of the company and to overcome this loss we came with an idea what if the employee gets the desired/expected salary from the Company or Organization. In this Competitive world everyone has a higher expectation and goals. But we cannot randomly provide everyone their expected salary there should be a system which should measure the ability of the Employee for the Expected salary. We cannot decide the exact salary but we can predict it by using certain data sets. A prediction is an assumption about a future event.

Linear regression algorithm in machine learning is a supervised learning technique to approximate the mapping function to get the best predictions. The main goal of regression is the construction of an efficient model to predict the dependent attribute from a bunch of attribute variables. A regression problem is when the output value is real or a continuous value like salary.

In order to gain useful insights into the job recruitment, we compare different strategies and machine learning models. The methodology different phases like: Data collection, Data cleaning, Manual feature engineering, Data set description, Automatic feature selection, Model selection, Model training and validation, Model comparison. We are focusing to develop a system that will predict the salary based on different parameters used in company and abovementioned methodology phases. Some of the parameters we collected from company data are: Job Type: CFO, CEO, Senior, vice president, manager

1. Degree: Doctoral, Bachelors, Masters, High School

2. Years of Experience

3. Country

4. Type of Developer

5. Salary

Motivation –

Nowadays prediction engine has become so popular that they are generating accurate and affordable predictions just like a human, and being using industry to solve many of the problems. Predicting justified salary for employee is always being a challenging job for an employer. In this project I am proposing a salary prediction model with suitable algorithm using key features required to predict the salary of employee.

Many websites like glassdoor and indeed predict the salary of an employee according to the given attribute and they need to be precise while doing this. I have tried to implement most of the models to find the best and most precise value here, to get the best predicted value here.

Methodology –

1. I imported the libraries needed for its implementation.

Pandas, numpy, matplotlib

1. Read the file and check all the columns and what are its values.
2. Take into account only the important columns.

Working on the salary column-

Removing the null values from the salary column.

Convert the float values into integers.

We will also remove the null values of all the other columns.

Working on Experience –

We will only convert the string values into integers.

So more than 50 and less 1 year is converted to 50 and 0.5.

Working on Education Level –

We will remove the unnecessary values from the degrees and just remain with the-

1. Bachelors
2. masters
3. less than bachelors
4. post doctorate.
5. def clean\_education(x):
6. if 'Bachelor’s degree' in x:
7. return 'Bachelor’s degree'
8. if 'Master’s degree' in x:
9. return 'Master’s degree'
10. if 'Professional degree' in x or 'Other doctoral' in x:
11. return 'Post grad'
12. return 'Less than a Bachelors'

We will remove the users with all the other type of values.

Working on the Developer Type-

We will just take into account the prime type of developer-

1. Full stack dev
2. Back end dev
3. Front end dev
4. Mobile Dev
5. Game Dev
6. Data Scientist
7. def clean\_devtype(x):
8. if 'front-end' in x:
9. return 'front-end developer'
10. if 'back-end' in x:
11. return 'back-end developer'
12. if 'mobile' in x:
13. return 'mobile developer'
14. if 'academic' in x:
15. return 'academic researcher'
16. if 'game' in x:
17. return 'game developer'
18. if 'data' in x:
19. return 'data scientist'
20. if 'full-stack' in x:
21. return 'full-stack developer'

Working on Country –

I don’t want the model to get confused and so, I’ll take into the account the countries having more than 300 developers.

I remove all the other developers from the dataset.

def remove\_countries(counts,bar):

    counts\_map={}

    for i in range(len(counts)):

        if counts.values[i]>=bar:

            counts\_map[counts.index[i]]=counts.index[i]

        else:

            counts\_map[counts.index[i]]="other"

    return counts\_map

Removing Outliers –

We need to remove the outliers from all the countries, countries like United States of America have big billionaires which makes a lot of difference.

We plot a box plot for checking those outliers.

fig,ax=plt.subplots(1,1,figsize=(12,7))

df.boxplot("Salary",'Country',ax=ax)

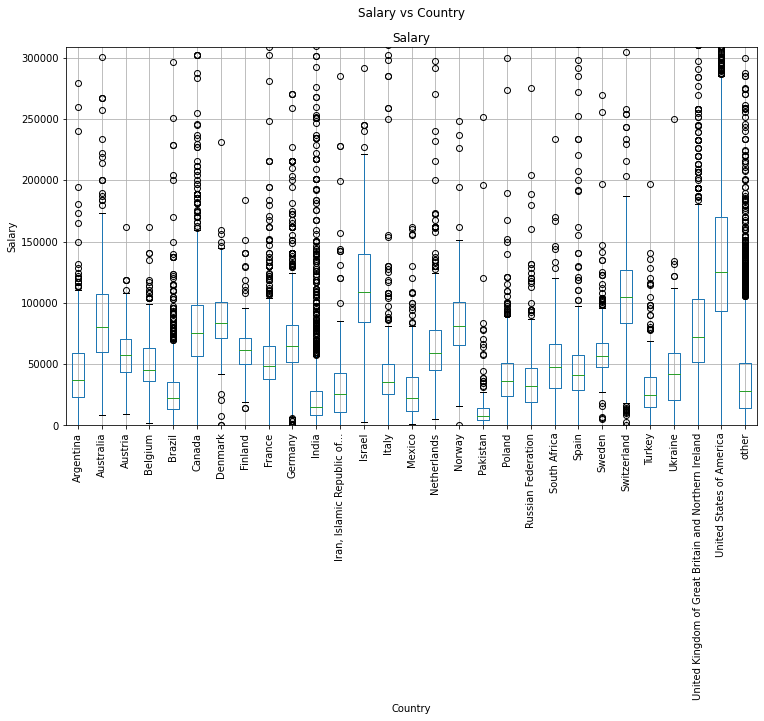
plt.suptitle("Salary vs Country")

plt.ylabel("Salary")

plt.xticks(rotation=90)

plt.ylim(0,308520)

plt.show()



we can see that we have a lot of outliers here, after changing the limits of salary various times, we arrive to the decision that to remove the outlier, we will limit our salaries, but it should still contain some higher and lower values, therefore max will be 250000 and lowest would be 10000, we will remove the other values.

Creation of Models-

List of models I created-

|  |  |
| --- | --- |
| Name | Mean Absolute Error |
| Linear Regression | 44035.000324853405 |
| Decision Tree Regressor | 26963.13126461602 |
| Random Forest Regressor | 27292.812864555683 |
| Grid Search | 31128.610701331247 |
| XG Boost | 29285.348949161387 |
| Light Gradient Boost | 42278.6891221509 |
| TensorFlow Keras | 44795.866636312276 |

To my surprise, Decision Tree Regressor is performing the best here.

But we will go ahead with it.

We will save our model in a pickle file.

Now we are going to deploy our model on Streamlit.

So, we will make app.py file, with predict\_page for prediction and explore for viewing some metrics in the form of graph.

Working on Streamlit-

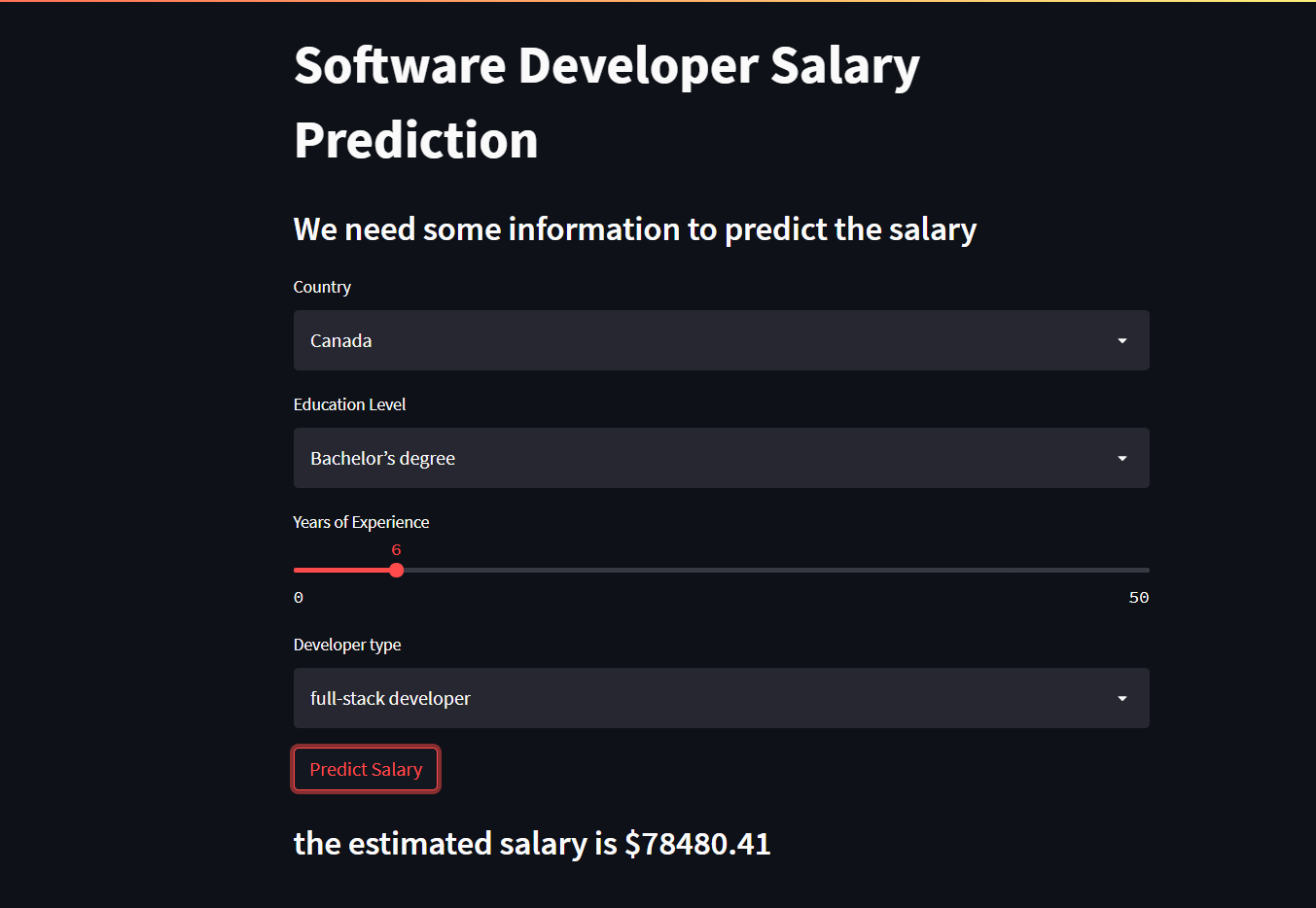
We will app.py and import the pages that is predict and explore.

In explore page- We will all our code and function to be executed.

And all the graphs we want to create.

In predict page- We will put the button and transformation.

We will print the salary with the help of streamlit function.



Flowchart-

References –

* Andrew NG course
* Hands on machine learning with Scikit-learn, Keras and TensorFlow
* <https://www.geeksforgeeks.org/machine-learning>
* dataset- Stackoverflow developer survey-

<http://insights.stackoverflow.com/survey>