Courseproject:Machine Learning(DataScience-PGC,Internshala Trainings)

Problem Statement:

Create a machine learning model which will help the company in determining the salary of newly hired employees using the given data.

Description:

TechWorks Consulting is a consulting company that specializes in providing top-notch IT talent to a variety of businesses. The company is known for its ability to quickly and efficiently find and place highly skilled IT professionals in a variety of positions.

The company was founded by a group of experienced IT professionals who saw a need for a better way to match talented individuals with the right job opportunities. They understood that the traditional job search process is often slow and ineffective, so they set out to create a company that could do things differently.

TechWorks Consulting has a unique approach to finding and placing IT professionals. They have a large and constantly growing database of qualified candidates, which they built through networking, referrals, and targeted recruiting efforts. They also have a team of experienced recruiters who are experts at identifying the best candidates for any given position.

Thanks to this approach, TechWorks Consulting has quickly become one of the most successful IT staffing companies in the country. Businesses of all sizes, from small startups to large corporations, come to rely on TechWorks Consulting to find the IT talent they need to grow and compete.

The company is particularly known for its ability to place large numbers of IT professionals in a short amount of time. They often work with businesses that are expanding rapidly and need to hire dozens or even hundreds of IT employees in a short period. TechWorks Consulting is able to handle these large-scale hiring projects with ease, thanks to its large database of candidates and experienced recruiting team.

TechWorks Consulting continues to set the standard for IT staffing and consulting. Businesses trust the company to provide them with the best IT talent in the industry, while IT professionals know that TechWorks Consulting is the best place to find the right job opportunity. The company is continuously expanding their services and domains to provide valuable services to their clients, making TechWorks Consulting one of the most sought-after consulting firms in the industry.

One of the key factors in TechWorks Consulting's success is its ability to provide fair and competitive compensation to its employees. The company takes multiple factors into account when determining an employee's salary,including the employee's experience, qualifications, and performance.

One important factor that TechWorks Consulting considers is the market rate for the specific job and skill set. The company has to make sure that the salary they offer is competitive with what other companies in the industry are offering for similar positions. This ensures that they are able to attract and retain top talent.

TechWorks Consulting also takes into account the specific skills and experience of the employee they are hiring. For example, an employee with more experience or specialized skills may be offered a higher salary than someone who is just starting out in the field.

They also consider the current market trends and the cost of living of the area where the employee will be working. So the salary packages are not fixed, they are flexible with the market trends and adjust accordingly.

Another important factor that TechWorks Consulting considers is the employee's performance. The company has performance evaluation systems in place to determine the employee's value and contributions to the company. The employee may be eligible for salary increases based on their performance and contributions.

In conclusion, TechWorks Consulting puts a lot of thought and effort into determining the salary of its employees. They take into account a variety of factors, including the market rate, the employee's experience and qualifications, and the employee's performance, to ensure that they are offering competitive and fair compensation.

Data

You are given employee data (**here**)as well as various other features that can be responsible for determining the employee's salary, such as the **college** an employee attends or the city from which the employee is coming, what the employee's previous CTC was, how much experience that employee has, and his academic record.

The data contains 8 columns:

College name: Colleges belong to three groups Tier1,Tier2 and Tier3 where tier1 college has the highest weightage.

City:It has 2 types of cities: metro and non metro cities convert this categorical data into numerical data such that 0 goes for non metro and 1 for metro cities.

Role: Manager and Executive

And other columns Like: Previous CTC,Previous Job Change,Graduation marks, Experience in Months and CTC.

Regression task

Machine learning, specifically regression, can be useful for TechWorks Consulting in determining or predicting the salary of an employee.

Regression is a machine learning technique that can be used to predict a continuous value, such as salary. TechWorks Consulting can use historical data about past employees, such as their qualifications, experience, performance, and salary, to train a regression model. Once the model is trained, it can be used to predict the salary of new employees based on their qualifications, experience, and other relevant factors.

The company can use a variety of features as inputs to the model such as education level, past job experience, certifications, location, etc. The model can then learn the relationship between these features and the salary of the employees.

One benefit of using machine learning to predict employee salary is that it can help to ensure that compensation is consistent and fair across different employees. By using a regression model, the company can make data-driven decisions about salary, rather than relying on subjective judgments or estimates.

Additionally, machine learning models are able to make predictions faster and more accurately than manual methods, especially with the increasing amount of data, this can help the company save time and make better predictions.

It's worth mentioning that machine learning is a tool and its results are only as good as the data it is trained on and the features you provide to the model, so TechWorks Consulting should make sure that their historical data is reliable, diverse and complete to make the best predictions. Also, having experts in machine learning and statistics can help to fine-tune the model to better predictions.

Statistics plays an important role in the field of machine learning, particularly in the development and evaluation of regression models.

In the process of developing a regression model for TechWorks Consulting, statisticians would first analyze the data to understand the distribution of the salary and other features. They would also investigate the correlation between different features and the target variable (salary) to understand which features are more important and relevant to the predictions.

Next, they would use statistical techniques to select a subset of the most important features to use as inputs to the regression model. This can be done through feature selection or feature engineering methods such as Lasso, Ridge, PCA and many others.

Once the model is trained, statisticians would use statistical methods to evaluate its performance. One common method used in evaluating regression models is Mean Squared Error (MSE), which measures the average difference between the predicted salary and the actual salary. A lower MSE indicates that the model is making more accurate predictions. Other metrics like R-Squared and AIC can also be used to evaluate the model.

Additionally, statisticians would also use statistical hypothesis testing to evaluate the significance of the model's coefficients and to make sure that the results are reliable and not just by chance.

In summary, statistics plays an important role in the process of developing and evaluating regression models for determining or predicting employee salary. TechWorks Consulting can benefit from having statisticians on their team to ensure that their models are accurate and reliable, and that the results are statistically significant.

Another step which plays a significant role for predicting the salary of an employee using regression is data pre-processing.

Data preprocessing is an important step in the machine learning pipeline, particularly when developing regression models for determining or predicting employee salary at TechWorks Consulting.

Data pre-processing involves cleaning and transforming the data to make it suitable for training a machine learning model. This can include tasks such as:

- Handling missing values: Some data points may be missing, statisticians would need to decide whether to remove the missing data or fill it in using techniques such as imputation.

- Handling outliers: outliers can greatly affect the model predictions and statisticians would need to identify and either remove them or transform them in a way that they would not impact the model's performance.
- Handling categorical variables: Many machine learning algorithms can only work with numerical data, so statisticians would need to convert categorical variables, such as education level, job titles, etc, into numerical values using techniques such as one-hot encoding.
- Normalization: Scaling the values of different features to the same range so that one feature does not dominate the others
- Feature selection: Choosing a subset of the most important features to use as inputs to the regression model

Data pre-processing is an iterative process and statisticians may need to test different pre-processing techniques and combinations of techniques to find the best approach for the specific data set and problem at hand.

By properly pre-processing the data, TechWorks Consulting can ensure that their regression model is trained on high-quality, accurate data and thus make more accurate predictions of employee salary. It can also prevent many issues that could arise later on such as overfitting, poor predictions and bias in the model.

Task:

As a data analyst at TechWorks Consulting, you are given the task of creating a machine learning model which will help them in determining the salary of newly hired employees using the given data.

You have to define your approach and then perform all the tasks required step by step to make the prediction, and then ultimately perform predictive analysis to predict the salary.

In data preparation you have to change the college field into numerical data type with the help of the tier of that college, similar thing you have to do with the city field and you also have to create dummy variables for “Role”.

Try to find out the outliers and missing values and clean the data further and after the data is ready create a predictive model to predict the salary.

You can choose any regression model but in the end you have to state your reason for choosing that model. You can try multiple models and then pick one with the best accuracy and also state the reason why this model performed better then the others and in what ways you can further improve the accuracy of the selected model.

You have to share the .ipynb file in which you will perform all of the required steps this file should also contain the answer of following question (you can use markdown option to answer these questions in same notebook)

1. Your views about the problem statement?
2. What will be your approach to solving this task?
3. What were the available ML model options you had to perform this task?
4. Which model’s performance is best and what could be the possible reason for that?
5. What steps can you take to improve this selected model’s performance even further?

Note:Your final deliverable is your Jupyter Notebook, which should contain all the preprocessing steps, visualizations, and the trained model. Additionally, it should also include the answers to the above 5 questions. You can use the markdown option of Jupyter Notebook to accomplish this

Evaluation Rubrics:

Q1: 10 Marks

- Attendance: 5 marks
- Correct explanation: 5 marks (additional)

Q2: 10 Marks

- Attendance: 5 marks
- Correct explanation: 5 marks (additional)

Q3: 10 Marks

- Attendance: 5 marks
- Correct explanation: 5 marks (additional)

Q4: 10 Marks

- Attendance: 5 marks
- Correct explanation: 5 marks (additional)

Q5: 10 Marks

- Attendance: 5 marks
- Correct explanation: 5 marks (additional)

Proper text representation of each step: 10 Marks

- Quality-dependent marks (ranging from 0 to 4 additional marks)

Code supported by proper comments: 5 Marks

- Quality-dependent marks (ranging from 0 to 2 additional marks)

Steps required before training the model: 5 Marks

Visualizations used in data preprocessing: 10 Marks

- Quality-dependent marks (ranging from 0 to 5 additional marks)

Working model: 20 Marks

It's important to note that the additional marks for quality are subjective and will vary based on the quality and accuracy of the work presented. The range mentioned is from 0 to the maximum additional marks achievable for each category.