International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Student Placement Package Prediction by Regression Analysis

Vaibhav Srivastava[1], *Saksham Kaushik*[2], *Ujjwal Raj*[3]

[1,2,3] School of Computing Science and Engineering, Galgotias University, Greater Noida (U.P.)

Doi: https://doi.org/10.55248/gengpi.5.0124.0345

ABSTRACT—

The "Student Placement Package Prediction Model" project aims to build a powerful model that is able to predict placement package of the student that is going to sit in placement exam based on their skills, academics performance, and extracurricular activities. The project uses Regression analysis to get the quantitative analysis of relationship between the input variables and the anticipated placement package.

The main aim of this project is to analyse the prediction that will help the students to decide their future career choices . By utilizing old placement data and the student regression model the system will provide accurate prediction aiding student to set realistic expectation in their career choices.

The methodology involve many rigorous model evaluation, employing metrics such as Mean Absolute Error (MAE), Mean Square Error (MSE), and R-Squared (R2) these all model ensures the model's accuracy, and effectiveness.

This project not only help student to decide their career choices but also it help institutions to decide and build their curriculum to help student to grow their learning and help them in their skill to stand out in the placement. This model assist in a more efficient, transparent, and mutually beneficial placement process.

Keywords—MAE(Mean Absolute Error), MSE(Mean Square Error), R2(R-Squared Error)

Introduction

In the pursuit of higher education, the education system has become elaborately inter wined with the aspiration for promising career prospects.

[1]By determining the placement packages has gained considerable attention within the system of education and career development.

Analyzing and accurately forecasting the placement package is not only important for educational institute but also very important to student who are preparing for the higher packages and also for those who wants to keep their skills and knowledge in check.

This study not only aims to determine the placement package but also work as the mirror for the institutions and the students, thereby providing the foundation for informed decision making and strategic educational planning.

Literature Survey

Introduction

The importance of predicting placement packages for students taking placement exams based on their skills, academic performance and extracurricular activities is gaining importance. Literature survey tells us about existing research and methodologies in the field, aiming to build a comprehensive understanding of the context and methodologies used in predicting student placement packages.

Educational Performance

While academic studies are important but some student start to focus mainly on the academic studies that make them lose the understanding of beyond knowledge. However, the literature emphasizes the need for a paradigm shift, advocating for holistic preparation the extends beyond academic brilliance.

Holistic Preparation Beyond Academics

Today most of the companies seeks students that have more practical knowledge than the students who have only knowledge of academics. This doesn't mean that the companies only prefer practical knowledge seeker but they want the student who have their knowledge balanced in both the field.

Process used to check the Performance

Before the emerge of A.I. most of the student check there performance by comparing it to their ideal they guess weather their performance is good or they are just doing without gaining but it lacks much accuracy compared to now as today students have papers to solve guider to guide them and much more.

The Signification Of Internship And Experiential Learning

Internship can also be called as the implementation of your skills in the real world application. They provide the platform where a student can apply all his knowledge to gain experience and knowledge. It encourages students to approach experiential knowledge as the important part of their carrier journey, driven by the genuine proactive learning mindset.

Tools and technology used in this model

This model is made by the help of regression analysis. As we all know regression analysis is used for the prediction and also we have used three types of regression model : Linear regression model[2], Multi linear regression model, artificial neural networks these all model gives predicted value making the students to understand whether they are learning correctly or not.

Technology means small errors

Errors are the fundamental part of the technology that every developer wants to remove but not in every case like if we know that the predicted of the model is somewhat less accurate then it means predicted value may come true if the student environment change. This is similar to whether today it will rain or not, now there is chance of half but what if prediction is made for the rain then it means two by three chances are there for the rain and if the prediction become wrong that doesn't mean that the overall situation is worst like in this model prediction is somewhat wrong student can change that prediction as this is the model that help to provide the prediction or idea of there package that is totally depends on the student.

Conclusion

The literature survey highlights the multifaceted nature of predicting student placement packages. The Importance of accurate predictions for students career decisions and institutional planning. The synthesis of diverse studies provides a comprehensive perspective on the current state of placement prediction research and sets the stage for the advancements proposed in the project.

Ease of Use

User Interface

The GUI provided in this have clear buttons for loading a CSV file and running the analysis. User gets the feedback about the action they need to take. If CSV file is not loaded user is prevented via alert from running the analysis.

Navigation and Interaction

The GUI provide easy navigation feature like the uploading of csv file and interaction of the load csv button with the assessing your file in computer and doing the analysis and providing the result in nice manner.

Training and On boarding

The csv file in model trains with 80 percent of the data user provide and then the test the trained data with the remaining 20 percent of the data. In the model there is a clear and easy use of resources.

Analysis Results And Code Structure

Results are presented in a clear and structured manner. Error valued are displayed for each regression models that are used and also errors like mean square error, mean absolute error and r-square error. Some Python Libraries are used to graphically represent the results using bar charts with different colors for model comparison. Code is organized well so that it can be read, maintain and can be improve.

Model Design

Machine learning models that are used in this project are predefined models like Linear Regression, Multiple Linear Regression and Neural Network so that it can simply the user experience and be understood more easily. Mat-plot Library is used to represent the graph between models and errors for a unified user experience.

Areas of Improvement

To enhance the user experience by providing more interactive interface and also less error with a beneficial result.

Continuous Improvement Through User Feedback

User feedback on error handling is actively sought and incorporated into iterative improvements. The system evolves based on the user insights, ensuring that the error handling process aligns with user expectations and experiences.

User Education

The system takes a proactive approach to user education by providing the guide and the relevant documentation to prevent errors. Pop -up messages, or contextual help s are strategically employed to the guide users in making informed decisions.

User Satisfaction Metrics

User satisfaction is a critical aspect of evaluating the rise or the downfall of the system , methodology, or tool. We will various surveys to satisfy the user to enhance the model.

User comments highlights the importance of the customized options or not `and that indicates a desire for more personalized experience without compromising the nature of the system.

Problem And Purpose

The problem that exists in this system is related to the prediction of the package that it will present as the related csv file contain all the related data about the placement that was held before. This projects aims to give maximum effectiveness of the linear regression, multiple linear regression and artificial neural networks mode in predicting package.

Purpose

User are allowed to load the csv file so that if a student have any other data that the user wants to train he/she is free to use.

The main purpose of this project is to load the placement data that the user have in a definite format so that it can be trained and provide necessary prediction and also helps in many ways.

The other purpose is to help students to adjust their study and university professor and teacher to adjust the curriculum that is going on.

It also give the basic idea of machine learning and python and regression analysis.

Learning about models that are used in this projects

Methodology

The methodology that involve in this project is the approaches that are used to predict the package. The approach used in this is regression analysis. Regression analysis is analytic method used to examine the relationship between the dependant variable and more than one independent variable. The main goal is to understand the relation between variable, strength and the nature of the variable.

Regression models used

Linear Regression Model: Linear Regression Model is used to understand the relation between a dependant variable and one and more independent variable by using linear equation to observe data.[3]

The general form of linear regression model is:

$Y=\beta0+\beta1X+\epsilon$

Where:

- Y is dependent variable
- X is independent variable
- $\beta0$ is the intercept (the value of Y when X is 0).
- $\beta1$ is the slope (the change in Y for a one-unit change in X).
- ϵ is the error term, representing unobserved factors that affect Y but are not included in the model.

Advantages of Linear Regression

- Linear regression technique is a very manageable algorithm that can give optimized solution. It can be implemented easily and effectively on low computational capacity system as compared other complex algorithms.
- The mathematical equation of the linear regression can be understand very easily and its time complexity is much lower as compared to other algorithm.

Challenges of Linear Regression

- Linear regression may stick to under fitting, a situation in which machine learning models fails to encapsulate the data of interest properly.

- Most of the naturally occurring phenomena are non-linear therefore linear regression technique fails to fit complex data sets properly because it assumes that there exists a linear relationship among the input and output variables.

Multiple Linear Regression Model: Multi linear extends the concept of linear regression by adding more independent variable to predict a dependent variable in more specific or accurate way.[4] It is useful when the linear regression can not catch up the relationship between the dependant variable and the predictors.

The general form of Multi linear Regression model is expressed as:

$Y=\beta 0+\beta 1X1+\beta 2X2+\ldots+\beta nXn+\epsilon$

Where:

- Y is the dependent variable.
- X1,X2,…,Xn are the independent variables.
- $\beta 0$ Is the intercept (the value of Y when all X variables are 0)
- $\beta 1,\beta 2,\ldots,\beta n$ Are the slopes (representing the change in y for a one- unit change in the corresponding X variable)
- ϵ Is the error term , according for the unobserved factors impacting Y not included in the model.

Advantages Of Multi Linear Regression

- It handles situation where the relationship between the dependent variable and multiple predictors is more complex than a simple linear equation.
- Incorporating multiple predictors often leads to improved accuracy in predicting the dependent variable.
- Allows for the assessments of how different variables interact with each other in influencing the dependent variable.

Challenges of Multi Linear Regression

- The presence of high correlation among independent variables of individual predictors
- Including too many independent variables may lead to over fitting, where the model performs well on the training data but poorly on new data.
- Multi linear regression assumes the quality and accuracy of the data used for training. Outliers or errors in the data can impact the model's performance.

Artificial Neural Network Model: ANN is employed to model complex relationships between variables, inspired by the human brain's neural structure.[5][6]They are systems that are able to modify their internal structure in relation to a function objective.

The general architecture of an artificial neural network involves:

- **Input Layer :** Receives input variables.
- **Hidden Layer : Process** information through weighted connections.
- **Output Layer :** Produces the final output.

ANN have high pattern recognition-like abilities, which are needed for pattern recognition and decision-making that are robust classifiers with the ability to generalize and make decisions from large and somewhat fuzzy input data.

Advantages Of ANN

- The main advantages of ANN is works with incomplete data and then produce the output. The output may be lost that is totally depends upon the importance of the missing information.
- In order for ANN to be able to learn, it is necessary to determine the examples the examples and to teach the network according to the determined output by showing these examples to the network.
- Artificial neural networks have numerical strength that can be performed more than one job at the same time

Disadvantages Of ANN

- Need for large amounts of labeled training data. Lack of transparency in decision-making.

- Computationally intensive and resource- consuming. Sensitivity to input data quality and prepossessing
- Potential over fitting without proper regularization. Complexity and difficulty in model tuning

Errors Used:

Mean Squared Error(MSE): Mean Square Error is the average of the squared differences between the predicted values and the actual values.[7]

General formula of the mean squared error is as follows:

$MSE=1/n\sum^{n}_{i=1}(Yi-Y\hat{}i)^{2}$

Where:

- N is the number of data points.
- Y_i is the actual value of the dependent variable for the i-th observation
- Y^I is the predicted value of the dependent variable for the I- th observation.

Advantages and challenges of Mean Square Error:

MSE value is always positive and its value shows the quality of the model. Lower value shows that the model is performing best and higher value shows model is not performing best. MSE is a continuous function making it compatible with optimizing technique. When optimizing problem using MSE as a loss function or continuous function this shows the model parameters that minimize the average squared difference between the predicted value and the actual value. MSE assumes that the errors are normally distributed and have constant variance. If these assumptions are violated, it may not provide accurate measure of model performance.

Mean Absolute Error: Mean Absolute Error is the average absolute differences between the predicted and actual value. It provides a straightforward representation of the average magnitude of the errors in predictions, making it easy to interpret.

General formula of the mean absolute error is as follows:

$MAE=1/n\sum^{n}_{i=1}|Yi-Y\hat{}i|$

Where:

- N is the total number of observations.
- Yi represents the actual values.
- Y^i represents the predicted values.

Advantages and challenges of Mean Absolute Error:

MAE is easy to understand and interpret since it represents the average absolute difference between predicted and actual values. The MAE value is the average of the absolute differences.Ma is beneficial when considering errors without the direction of overestimation and underestimation. MAE assumes even loss, meaning overestimating and underestimating the target variable have the same impact.

R-Square Error: R square allows us to measure the proportion of the variance in the dependent variable and the independent variables. [8]The r- square values ranges from 0 to 1 with 0 indicates that the model does not explain any variance and 1 indicates a perfect explanation of the variance.

General formula of R-Square is as follow:

$R2$=1−(Sum of Squares of Residuals/

Total Sum of Squares)

Where:

- Sum of Squares of Residuals : The sum of the squared differences between the actual values and the predicted values
- Total Sum Of Square : The sum of the squared differences between the actual values and the mean of the actual values.

Advantages and challenges of R-Square Error:

R-Square provides a numerical measure of how well the regression model fits the observed data. It provide the facility to compare different models. R Square helps identify the importance of different independent variables and the dependent variable. R-Square does not indicate the causation. Even if the model fits the data well, it doesn’t mean that the changes in the independent variables cause any change in dependent variable.

Conclusion

The performance evaluation of the student placement package prediction model demonstrates its effectiveness in forecasting placement packages based on the comprehensive set of independent variable. Through rigorous testing and validation,the model exhibits promising accuracy and robustness, providing valuable insights into potential placement outcomes for students.

The precision and recall metrics showcase the model's ability to correctly identify positive placement cases while minimizing false positives. Overall, the performance evaluation underscores the model's reliability and utility in supporting decision-making processes related to student placements.

Advantages Over Existing Technologies:

This model leverages a rich set of attributes leading to enhance predictive accuracy compared to traditional models. This model reflect a more inclusive approach to placement predictions. The model's adaptability to changing educational landscapes and industry requirements ensures relevance in dynamic industry markets.

Future Scope

As we look forward for the enhancement and expansion of the Student Placement Package Prediction Model are visualized.

The integration of advanced artificial intelligence technique, such as deep learning, to capture correlations in large-scale datasets. Improving the model's user interface to enhance user experience, making it more accessible for the educational institutions, career counselors, and students.

In conclusion, the student placement package prediction model not only shows strong performance but also lays the groundwork for continuous improvement and adaptability to meet the evolving needs of industry.

Acknowledgment

The authors are thankful to Prof. Saurabh Singh. Department of Computing Science And Engineering, Galgotias University for his valuable suggestions and guidance while preparing this paper.

References

Marko Sarstedt, Erik mooi, "Regression Analysis", Aconcise Guide to Market Research (pp. 193-223), March 2014

Gibbs Y. Kanyongo, Janine certo, and Brown I. Launcelot "Using regression analysis to establish the relationship between home environment and reading achievement: A case of Zimbabwe" by GY Kanyongo, 2006.

Muhammad Ahmad Iqbal, "Application of Regression Techniques with their Advantages and Disadvantages", Article 4, UET Lahore.

Gülden Kaya Uyanık, and Neşe .Güler, "A Study on mutiple linear Regression Analysis", Social and Behavioral Sciences 106 (2013) 234 - 240 [4th International conference on new Horizon in ediucation].

Enzo Grossi, and Massimo Buscema, "Introduction to artificial Neural Networks", European Journal of Gastroenterology & Hepatology 19(12):1046-54,January 2008.

Maad M. Mijwil,"Artificial neural networks Advantages and Disadvantages",Mesopotamian journal of Big Data 2021:29-31.

Alexei Botchkarev,"Performance Metrics (Error Measures) inn machine Learning Regression, Forecasting and Prognostics: Properties and Typology",2018.

Ozili, Peterson K,"The acceptable R-square in empirical modelling for social science research", MPRA Paper No. 115769, 26 Dec 2022.