Engineering Graduate Salary Prediction Using Principal Component Analysis

Abha Marathe[a], Arjun Thakur[b], Shubham Shah[b], Shruti Singh[b], Akash Sinha[b], Vishal Sirvi[b], Shreyansh Srivastava[b]

[a] Department of Electronics and Telecommunication Engineering, Vishwakarma Institute of Technology, Pune.

[b]Department of Computer Engineering, Vishwakarma Institute of Technology, Pune.

***Abstract* — The goal of this study is to establish quantitative methodologies for estimating an engineering graduate student's salary. What influences the income and work opportunities for engineers immediately after graduation is a pertinent topic. Several aspects like college grades, applicant talents, college proximity to industrial centres, specialization and market requirements for certain sectors have a significant influence . A comparison of different regression models, including Multiple Linear Regression (MLR), XGBoost for regression, Support Vector Regression (SVR), Random Forest for regression, and Decision Trees for regression, , is presented, taking all of the factors into account. After evaluating the performance of the respective regression algorithms, it was inferenced that the SVR provides highest R-squared value i.e., 0.935, followed by XGBoost with 0.91, random forest with 0.82, MLR with 0.59 and decision trees with 0.50 as the corresponding R-squared values.**

Keywords *— Salary Prediction, Principal Component Analysis, Random Forest , Multiple Linear Regression,Support Vector Regression, Decision Trees.*

INTRODUCTION

Major reason for undergraduate students to often worry is the salary which they will receive after graduation. Students keep applying for internships and keep doing various courses and projects to improve their skills thus increasing their chances of getting the expected salary[12]. In the current era of competitiveness, each individual has a higher expectation but it is not feasible to provide an amount that aligns with the expected salary for each individual. There should be a system which should measure the ability of a graduated student and mention the expected salary aligning to the market trends. We can't anticipate the precise salary, but we can use specific data sets to estimate it. A forecast is an educated assumption about what will occur in the future. In this paper, the main aim is to predict the salary of an engineering graduate student so that a student can understand what salary can be expected based on corresponding qualification and hard work. For developing this system, a dataset of Engineering graduate salary prediction has been used. Different algorithms have been applied on it like the Linear regression algorithm, Random Forest, Decision Tree, Support Vector Machine (SVM) Regression and Xgboost. The linear regression algorithm, a supervised learning strategy is used to approximate the mapping function and produce the best prediction. Random Forest Regression employs ensemble learning methods to aggregate predictions from multiple machine learning algorithms and produce a more accurate forecast when compared to a single model. In the shape of a tree structure, decision trees construct regression or classification models. [1] A decision tree having decision nodes and leaf nodes as an end result is built while the dataset is reduced into smaller chunks. Support Vector Machines analyze data used for classification and regression analysis. The straight line required to fit the data is referred to as the hyperplane in Support Vector Regression. XGBoost is a fast gradient boosting solution that can be used for regression predictive modeling. The primary goal of regression is to create a model that can predict the dependent attribute from a set of attribute variables. A regression problem occurs when the output value is a real or continuous variable, such as a salary.

LITERATURE REVIEW

Machine learning techniques are employed in the study [2] to automate and build a proposed salary prediction model. On the basis of a few important features, the proposed model can forecast salary.A raw dataset is fitted into decision models such as decision trees and ensembles models in the suggested approach

The process of predictive analysis begins with importing the necessary libraries and then performing data preprocessing over the dataset which is to be trained thereafter by using Logistic Regression and SVM[3].

A salary prediction system based on data mining techniques is presented in this research. The system compares the years of experience of the employee with the annual compensation[4].

Decision tree classifiers are sensitive to the dataset on which they are trained, according to a comparative examination of classification approaches. If the training dataset is modified, the resulting tree, as well as the accuracy of the prediction, may be drastically different. This problem is addressed[5] using the Random forest classifier which gives better performance than decision trees.

Linear regression performs a task in which it searches for

a linear relationship between input x and output y. In polynomial regression, the link between the unbiased variable x and the dependent variable y is handled as an nth degree polynomial[6].

Performance score, Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R2 Score has been used to evaluate the performance of various regression approaches such as Random Forest Regression, Multilinear Regression, and Decision Tree Regression.[7] .Multilinear Regression has a performance score of 0.6443, Random Forest has a score of 0.9643, and Decision Tree has a score of 0.616.

In the next paper, the authors have collected a real time dataset for the previous two years placement record from Ramrao Adik Institute of Technology (RAIT) training and placement office from Navi Mumbai, India having more than 200 records. For the research findings the authors have applied popular algorithms like Random Forest Regression, Multilinear Regression and Decision Tree Regression. The findings are as follows; The Multilinear Regression model performed the best with R squared value 0.6153 and 0.4289 for regular and diploma models respectively followed by Random Forest Regression and Decision Tree Regression[8].

Principal Component Analysis, is generally implemented when the input feature dimensions are high and is particularly useful for data processing where multicollinearity exists between the variables. The PCA application maximizes the variance of the original data set[9].

The main goal of the study is to anticipate efficient candidates with their skills and expectations of salary to ensure a convenient and cost effective process[10] .This model is built utilizing multiple statistical measures on feature selection and various machine learning algorithms to predict recruiting prospects based on certain key quantitative and qualitative factors such as age, gender, work experience, current wage and salary rises, and so on. These results will aid in predicting which candidates will eventually join.

METHODOLOGY

A. Dataset Description

Dataset Selection plays a significant role in what type of algorithm should be used where it can be supervised, unsupervised or semi-supervised as well as the quantity of valid records and attributes must be taken into account when selecting or importing datasets. The supervised technique is used in this paper on a dataset from Kaggle[11] consisting of 2998 records and 34 attributes out of which 19 appropriate attributes have been selected. The number of records and attributes chosen from the dataset were substantial enough to allow for the development of an efficient model and forecast a conveniently wider spectrum of options. The dataset is saved in CSV format and its dataset's qualities are logical and competent, allowing for a precise and efficient forecast of an engineering graduate's salary.

Name of the Attribute	Type	Categories
CollegeGPA	num	73.8 65 61.9 80.4 64.3 ...
GraduationYear	int	2013 2014 2011 2013 2012
English	int	650 440 485 675 575
Logical	int	665 435 475 620 495 595
Quant	int	810 210 505 635 365 620
Domain	num	0.694 0.342 0.825 0.99 0.278 ...
ComputerProgramming	int	485 365 -1 655 315 455 -1 465 525 385 ...
ElectronicsAndSemicon	int	366 -1 400 -1 -1 300 -1 -1 -1 -1 ...
ComputerScience	int	-1 -1 -1 -1 -1 -1 -1 -1 438 407 ...
MechanicalEngg	int	-1 -1 -1 -1 -1 -1 469 -1 -1 -1 ...
ElectricalEngg	int	-1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
TelecomEngg	int	-1 -1 260 -1 -1 313 -1 -1 -1 -1 ...
CivilEngg	int	-1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
conscientiousness	num	-0.159 1.134 0.51 -0.446 ...
agreeableness	num	0.3789 0.0459 -0.7473 ...
extraversion	num	1.24 1.24 1.543 0.317 -1.07 ...
neuroticism	num	0.1459 0.2727 0.0622 ...
openess_to_experience	num	0.289 -0.286 0.48 0.186 ...
Salary	int	445000, 110000, 255000 ….

B. Data Preprocessing

Data Cleaning

The data cleaning method entails finding the data, extracting it, cleaning it, and integrating it into a dataset that can be analyzed as per requirements. On visualizing the data, outliers were discovered in the salary column of the dataset which had to be eliminated since they decrease the correlation by scattering the data and causing r to approach zero. As a result, the salary column outliers were removed using their index positions to enhance the r2 value. In the bulk of the dataset, -1 values were

likewise a key source of interference since they resemble empty entries hence had to be eliminated too. As part of the data cleansing process, all -1 values in the selected attributes were converted to NA values, which were then subsequently replaced by the mean values of the corresponding columns.[16]

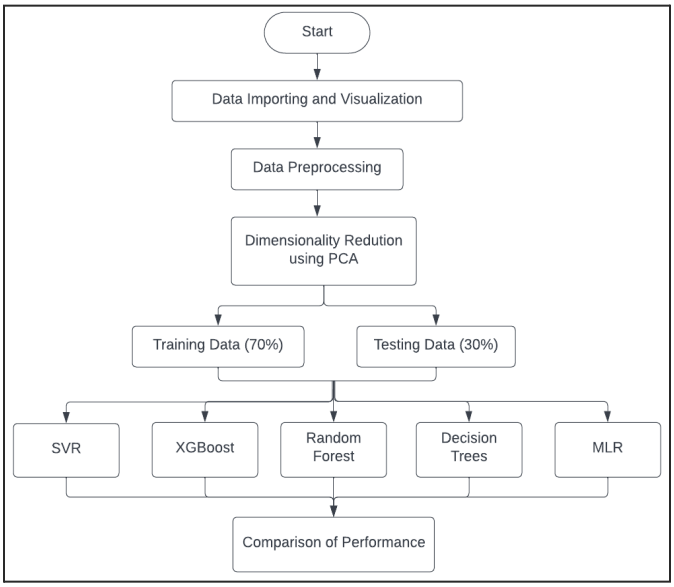

Fig 1. Workflow

Principal Component Analysis (PCA)

Principal components analysis (PCA) is used for the dimension reduction of data and helps in determining the most significant features in the dataset through a statistical approach. It ensures minimum loss of information and increases the interpretability of the data[18]. Principal components are the orthogonal projections which map a data having high dimension onto a low dimensional space. Thus, a principal component represents a line that encapsulates the majority of the data variance such that the first principal component captures most of the variance and for the i^{th} principal component contribution decreases as the i value increases.[15]

Firstly, The raw data matrix X_R with 19 features is centered to the mean, then the centered data matrix X is used to compute the covariance matrix C which is given as,

$$C = \frac{X^T X}{(n-1)}$$

where, C denotes the covariance matrix, X is the mean centered data matrix, X^T is the transpose of X and n represents the total number of samples in the data[10].

The covariance matrix is then used to compute the eigenvectors and corresponding eigenvalues λ_i i.e. eigendecomposition of the covariance matrix is carried out. As the matrix C is symmetric and positive semidefinite, so it is further diagonalized as follows,

$$C = V S V^T$$

where, V represents an orthogonal matrix consisting of eigenvectors and S denotes a diagonal matrix with non-negative eigenvalues λ_i as the diagonal elements, in decreasing order.

Thus, the columns of the matrix $P = XV$ represents the principal components. Out of the 19 principal components 15 are selected whose cumulative value of variance explained for the respective PCs is less than the threshold value 0.93, the cumulative variance explained is given as,

$$CVE\ (\%) = \frac{\sum_{k=1}^{i} \lambda_k}{\sum_{j=1}^{n} \lambda_j} * 100$$

where, λ_i denotes the i^{th} eigenvalue. The selection of the first few significant principal components as per specified threshold value of cumulative explained variance, ensures stability in estimation of salary [14]. Thus, the salary column is bound to the 15 selected principal components and they are splitted into training and testing dataset in the ratio of 70:30.

C. Regression Modeling

Linear regression (LR)

Regression is basically a statistical tool which is most commonly used in understanding and defining the relation between two variables. In regression there are two variables involved, the predictor variable and the response variable. The predictor variable is the one which is collected from a specific survey or an experiment and it is used to derive another variable. The variable which is to be derived with the help of the predictor variable is called the response variable.

In simple linear regression, the relation between predictor and response variable is defined by a mathematical equation of straight line where the powers of the variables is limited to 1.

$$y = ax + b$$

where y denotes the response variable, x denotes the predictor variable and a, b are the respective coefficients.

For a non linear type of relationship among the corresponding variables, polynomial regression can be used. In polynomial regression the exponent of the variable in the equation is not equal to 1 which generates the curve.

Multiple linear regression is another way to identify the relationship among multiple variables at the same time.[13]

After all the preprocessing and applying PCA to predictor variables for the given data set, when linear regression was applied to the data the R-squared value was found to be 0.5875523 and root mean squared error (RMSE) value was 104348.5

Support Vector Regression (SVR)

Support vector machines prove to be a competent algorithm that can significantly explain the non-linearity

of the data. Some of the crucial hyperparameters involved in SVR are hyperplane and kernel. Decision boundaries for forecasting continuous outputs are represented by the hyperplanes. The data points closest to the hyperplane, lying on each side of it are called Support Vectors. These support vectors are used to construct the necessary line depicting the algorithm's predicted outcome. The kernel is a mathematical function which is responsible for mapping non-linear data to linear data with high dimensional space.
Radial kernel is the default SVM kernel, which is given by the mathematical equation,

$$K(x,\, x') \;=\; e^{-\gamma\, ||x - x'||^2}$$

where, γ denotes a scalar that signifies the influence of the training data points and $||x - x'||^2$ represents the square of euclidean distance between two column vectors. When SVR is applied to the data obtained after PCA, it gives R-squared value of 0.9350 and root mean squared error (RMSE) value as 29929.55

Extreme Gradient Boosting (XGBoost)
XGBoost follows Boosting techniques in Ensemble Learning where the main aim is to correct the errors made by previously applied models in multiple iterations in which specific weights are added to the models. XGBoost [6] provides enhanced performance and features which include Auto-Pruning, inbuilt Cross-Validation and regularizations which reduce overfitting. Some parameters associated with the model are ƛ , ϒ and S.S which are used to signify the regularization,auto pruning and similarity score parameters respectively. eta, termed as learning rate, has been set to 0.25. n-rounds when set to 959 gives the best iteration where the max_depth is set to 6 and objective is reg:squarederror.
Similarity Score(S.S) is used for the gains and is calculated by the equation.

$$\text{S.S.} = \frac{(S.R.)^2}{(N + \lambda)}$$

where S.R denotes the sum of residuals and N denotes the number of residuals.
The S.S score is instrumental during the decision tree formation and corresponding S.S score for each leaf is again calculated for the Gain which is given by the difference of S.S of the branch before and after splitting. XGBoost, when applied after preprocessing and PCA application gives the R-squared value as 0.9138152 and root mean squared error (RMSE) value as 34827.99.

Random Forest Regression
Random Forest Regression is used for regression which uses ensemble learning methods. The ensemble learning method basically aggregates and assembles predictions from several machine learning algorithms and hence produces a more accurate forecast than a single model. It is like a combination of multiple Decision Trees. The trees run in a straight line with no interaction. [5] During training, a Random Forest constructs many decision trees and outputs the mean of the classes as the prediction of all the trees. [13] After applying the random forest model on the given dataset after preprocessing and principal component analysis, the following results are obtained. The mean absolute error (MAE) is found to be 32130.12. The mean squared error (MSE) is obtained as 1790945393 and the root mean squared error as 42319.56. The R squared value comes out to be 0.8275663.

Decision Tree Regression
In the decision tree approach, a regression model is constructed in the shape of a tree structure. It gradually cuts a dataset into smaller sections and develops an associated decision tree. A tree containing decision nodes and leaf nodes is the end result. [2] A decision node can have two or more branches each of which represents a value for the attribute being checked. A choice on the numerical aim is represented as a leaf node The topmost decision node present in a tree which shows the best predictor is represented by the root node.Decision trees are able to handle both category and numerical data. After applying the decision tree model on the given dataset after preprocessing and principal component analysis, the following results are obtained. The mean absolute error (MAE) is found to be 60571.13. The mean squared error (MSE) is obtained as 5990339904 and the root mean squared error as 77397.29. The R squared value comes out to be 0.5008397.

RESULTS

The results obtained after application of various machine learning algorithms on the columns before and after preprocessing are depicted in the form of a line chart where the black line represents the actual values.

Before PCA:

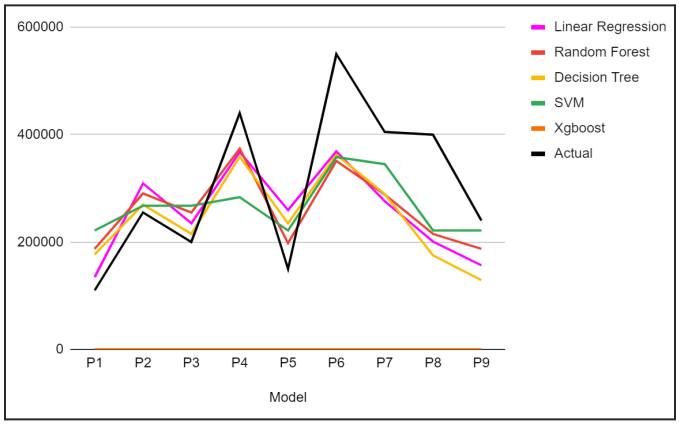

Fig 2. Line chart depicting actual and predicted values of all the models before PCA

From the above line chart it can be seen that none of the algorithms performed up to the mark and hence the PCA was applied in order to improve the results. The line chart is as follows:

After PCA:

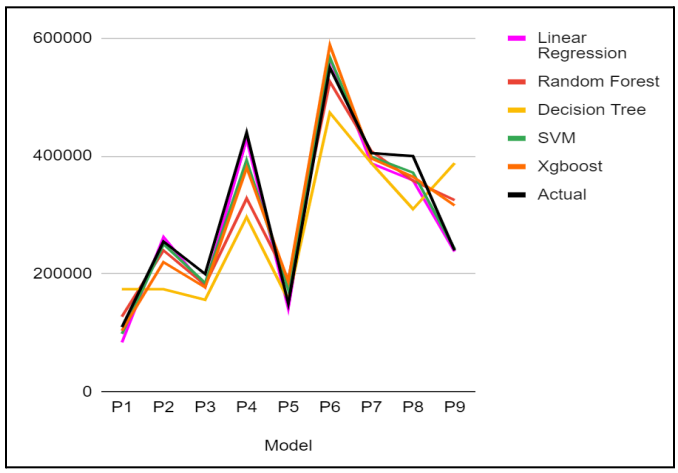

Fig 3. Line chart depicting actual and predicted values of all the models after PCA

The green line represents the SVM predictions. The green line is the closest to the black line. This indicates that SVM predictions are closer to the actual predictions when compared to other models' predictions. Hence the accuracy of prediction was increased after applying PCA . The results of machine learning algorithms on these modified columns are

Algorithm	MAE	RMSE	R-Square
Linear Regression	18997	104348.5	0.5875523
SVM	19405.87	29929.55	0.9350692
Xgboost	26151.12	34827.99	0.9138152
Random Forest	32130.12	42319.56	0.8275663
Decision Tree	60571.13	77397.29	0.5008397

Table 1. Performance Comparison after PCA

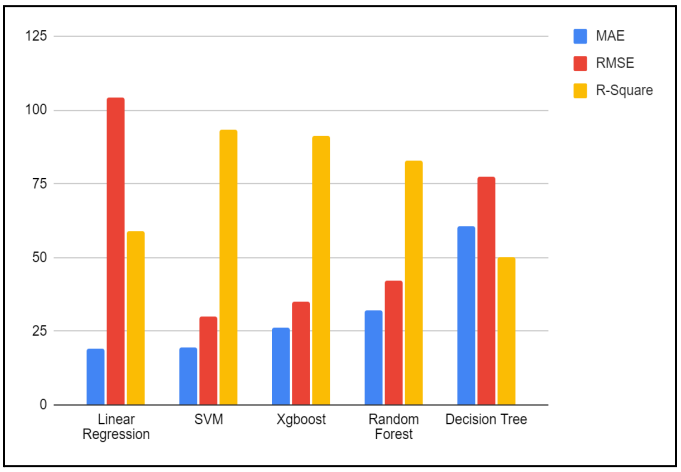

Fig 4. Bar chart depicting performance of different models after PCA

After comparing the results the best R-squared value of 0.9350692 was obtained by applying the Support Vector Machine(SVM) Regression. The following is the graph of the comparison:
In the above chart the values of respective columns have been scaled between a range from 0 to 125 for a proper view of the graph representing comparison.

Conclusion

Various algorithms have been applied on the data so as to get the best possible results. Firstly the data was preprocessed by imputing null values with the mean values and all the unnecessary columns were removed. Then the correlation between the variables were found and those variables which are highly correlated were removed and then different algorithms were applied. However, the obtained results were not satisfactory.
Hence, finally the concept of Principal components analysis (PCA) was used and the principal components for the data were found. Total 19 principal components were obtained. Out of theses 19 principal components, 15 principal components were selected whose cumulative value of variance explained for the respective PCs is less than the threshold value 0.93
Then the algorithms were again applied on this modified data which in turned improved the results.
Among the five algorithms that were applied, the performance of Support Vector Machine (SVM) Regression was found to be the best with R-square value as 0.9350692 and root mean square error (RMSE) 29929.55
The performance order of different algorithms starting with the best results on our dataset is as follows:
Support Vector Machine(SVM), Xgboost, Random Forest, Linear Regression, Decision Tree.
Hence we conclude for the given dataset Support Vector Machine(SVM) and Xgboost are the best algorithms to apply.

References

1. Chen, Jingyi, Shuming Mao, and Qixuan Yuan. "Salary prediction using random forest with fundamental features." In Third International Conference on Electronics and Communication; Network and Computer Technology (ECNCT 2021), vol. 12167, pp. 491-498. SPIE, 2022.
2. Dutta, Sananda, Airiddha Halder, and Kousik Dasgupta. "Design of a novel Prediction Engine for predicting suitable salary for a job." In 2018 Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), pp. 275-279. IEEE, 2018.
3. Kavitha S, Varuna S and Ramya R, "A comparative analysis on linear regression and support vector regression," 2016 Online International Conference on Green Engineering

and Technologies (IC-GET), 2016, pp. 1-5, doi: 10.1109/GET.2016.7916627.
4. Navyashree, M., M. K. Navyashree, M. Neetu, G. R. Pooja, and Biradar Arun. "Salary prediction in IT job market." Int. Journal of Comp. Sci. and Engineering 7, no. 15 (2019).
5. Das, Sayan, Rupashri Barik, and Ayush Mukherjee. "Salary prediction using regression techniques." Proceedings of Industry Interactive Innovations in Science, Engineering & Technology (I3SET2K19) (2020).
6. S. Ghosh, A. Dasgupta and A. Swetapadma, "A Study on Support Vector Machine based Linear and Non-Linear Pattern Classification," 2019 International Conference on Intelligent Sustainable Systems (ICISS), 2019, pp. 24-28, doi: 10.1109/ISS1.2019.8908018.
7. Srivastava, Suyash, Deepanshu Sharma, and Priyanka Sharma. Comparing various Machine Learning Techniques for Predicting the Salary Status. No. 2625. EasyChair, 2020.
8. Kakade, Saili, Pooja Shirude, Snehal Patil, and Balwant J. Gorad. "Analyzing and Forecasting the Students Placement Package." Available at SSRN 3884634 (2021).
9. N. Vaswani, Y. Chi and T. Bouwmans, "Rethinking PCA for Modern Data Sets: Theory, Algorithms, and Applications [Scanning the Issue]," in Proceedings of the IEEE, vol. 106, no. 8, pp. 1274-1276, Aug. 2018, doi: 10.1109/JPROC.2018.2853498.
10. Saini, Bhavna, Ginika Mahajan, and Harish Sharma. "An Analytical Approach to Predict Employability Status of Students." In IOP Conference Series: Materials Science and Engineering, vol. 1099, no. 1, p. 012007. IOP Publishing, 2021.
11. Engineering Graduate Salary Prediction [Online] Available:https://www.kaggle.com/datasets/manishkc06/engineering-graduate-salary-prediction
12. More, Anuj, Amay Naik, and Sarita Rathod. "PREDICT-NATION Skills Based Salary Prediction for Freshers." Available at SSRN 3866758 (2021).
13. Manoj, Jyothi, and K. K. Suresh. "Forecast Model for Price of Gold: Multiple Linear Regression with Principal Component Analysis." Thailand Statistician 17, no. 1 (2019): 125-131.
14. Davino, C., R. Romano, and D. Vistocco. "Handling multicollinearity in quantile regression through the use of principal component regression." METRON (2022): 1-22.
15. Mor-Yosef, Liron, and Haim Avron. "Sketching for principal component regression." SIAM Journal on Matrix Analysis and Applications 40, no. 2 (2019): 454-485.
16. Pahwa, Ashish, and Deepali Kamthania. "Quantitative analysis of historical data for prediction of job salary in India-A case study." Journal of Statistics and Management Systems 22, no. 2 (2019): 187-198.