Predicting The Annual Salary of Employees From AMEO Data Set Using Sci-kit

Md Fahim Shahriar
Computer Science and Engineering
Independent University, Bangladesh
Dhaka, Bangladesh
2120552@iub.edu.bd

***Abstract*—This paper investigates the use of machine learning to predict the annual salary of engineering candidates in India. Three regression models—Linear Regression, Decision Tree, and Random Forest—were trained and evaluated using cross-validation. Despite its simplicity, Linear Regression outperformed other models. However, further refinement is necessary to improve predictive performance.**

***Index Terms*—AMEO, Sci-Kit, Label, Feature, Linear Regression, Decision Tree, Random Forest**

I. Introduction

In today's competitive job market, the ability to accurately predict the annual salary of engineering candidates is of paramount importance for employers and job seekers alike. With the increasing availability of large-scale datasets containing comprehensive information about candidates, coupled with advancements in machine learning techniques, it has become possible to construct robust predictive models for this purpose.

The dataset used in this study contains a wealth of information about engineering candidates in India. Specifically, it includes scores from the Aspiring Minds' AMCAT test, which assesses cognitive, domain, and personality skills, along with personal information such as gender and date of birth, pre-university information like high school grades and location, and university information including GPA, college major, and college reputation proxy. Additionally, demographic information is also included.

In this paper, we present the construction and evaluation of machine learning models using the scikit-learn library to predict the annual salary of engineering candidates based on their profile information. Feature selection techniques, including correlation matrix analysis and selectKbest were employed to identify the most relevant predictors for the model. It is to be noted that before running the aforementioned models, the training data and the testing data was cleaned by omitting empty values and then was pre-processed to rectify any shortcomings.

II. Dataset Description

The AMEO (Aspiring Minds' Employment Outcomes) 2015 is a collection of dataset which contains comprehensive description of engineering candidates in India and their employment details. The dataset in which the all of the testing was to prepare it were categorized as Training Data and Testing Data. The training data is used to train the model. The testing data is used to evaluate the accuracy of the trained algorithm. This dataset provides valuable insights into the factors influencing job placement and annual salary for engineering graduates.

The Features that are expanded upon in this dataset are as follows:

- AMCAT scores (cognitive, domain, and personality assessments).
- Personal information (gender, date of birth).
- Pre-university information (high school grades, high school location).
- Demographic information (location of college, candidates' permanent location).
- AMCAT scores (cognitive, domain, and personality assessments).

By Utilizing these aforementioned features we are trying to predict the target variable 'Annual Salary'.

III. Data Cleaning and Pre-processing Steps

The AMEO dataset underwent a rigorous cleaning process to ensure data integrity and to prepare it for the construction of machine learning models.

Firstly, Columns such as 'Salary', 'DOJ' (Date of Joining), 'DOL' (Date of Leaving), 'Designation', and 'JobCity' were removed from the dataset. Since these columns had no values in the testing data, they were considered unnecessary for analysis.

Then, Features like '10board', '12board', 'CollegeID', 'CollegeTier', 'CollegeCityID', 'CollegeCityTier', and 'CollegeState' were removed from the dataset. These features were deemed to have little predictive power regarding the annual salary.

After that, the 'ID' column was designated as the index column since it is unique in both datasets and is used as a primary identifier.

Missing values represented as 0 and -1 were imputed, particularly in the 'Graduation Year' and 'Domain' columns, after using the info method on both testing and training data set.

From the 'DOB' (Date of Birth) column, the birth year was extracted. Subsequently, the graduation age was calculated,

Identify applicable funding agency here. If none, delete this.

as it was deemed to be a potentially significant predictor for salary. Graduation age may indicate candidates who graduated earlier due to being prodigies or candidates who experienced delayed graduation due to dropping out.

The ’specializations’ column was mapped using a dictionary to ensure uniformity and consistency in both the training and testing data set.

After acknowledging the presence of outliers in the dataset, we proceeded with data preprocessing. Categorical columns were encoded using ’One Hot Encoding’ and ’Label Encoding’, while numerical columns were scaled using ’Standard Scaler’.

- Label Encoding: Categorical columns were initially encoded using ’Label Encoding’, where each category is assigned a unique integer. However, this approach assumes an order or hierarchy among categories, which may not always be appropriate. Therefore, for categorical columns with a large number of categories, we used ’One Hot Encoding’ to avoid such assumptions.
- One Hot Encoding: For categorical columns with a large number of categories, such as ’Degree’ and ’Specialization’, we utilized ’One Hot Encoding’. This technique creates a separate binary column for each category, ensuring that the model does not assume any ordinal relationship among the categories.

As such the following columns were processed:

- For the ’Gender’ column, we used ’LabelEncoder’ from scikit-learn.
- or the ’Degree’ and ’Specialization’ columns, we used the pandas ’get dummies’ method to perform ’One Hot Encoding’.
- Standard Scaler: Numerical columns were scaled using ’Standard Scaler’ to ensure that all numerical features contribute equally to the model.

After the pre-proccesing steps there were 17 categories left in both the training and testing data set which outlined in the table below.

Remainder Features
”Gender”
”10percentage”
”12percentage”
”Degree”
”Specialization”
”collegeGPA”
”English”
”Logical”
”Quant”
”Domain”
”conscientiousness”
”agreeableness”
”extraversion”
”nueroticism”
”openess$_{t}o_{e}xperience$”
”12GradAge”
”GradAge”

TABLE I
RemainingFeatures After Data Processing

IV. Model Description and Feature Selection

In order to identify the most relevant features for our predictive model, we employed two feature selection techniques: correlation matrix and SelectKBest method.

In **Correlation Matrix**, we began by examining the correlation between features and the target variable using a correlation matrix. Features with a correlation coefficient above a certain threshold were selected for further analysis. Initially, we experimented with two threshold values, 0.1 and 0.2, to determine the optimal set of features.

Out of the 17 categories that remained after the cleaning process, the correlation matrix identified 7 categories that exhibited the strongest correlation with the target variable.

In **SelectKBest Method**, to validate the results obtained from the correlation matrix, we employed the SelectKBest method from the scikit-learn library. This method selects the top k features based on their importance scores. We set k = 7 to match the number of features identified by the correlation matrix.

Using the SelectKBest method, and after necessary preprocessing steps with the help of scikit-learn library, we obtained the exact same 7 columns out of the initial 17 categories (in TABLEI) identified by the correlation matrix.

By combining the results of both techniques, we were able to select the most relevant features for our predictive model.

Feature Selection Method	Columns
Correlation Matrix (With Threashold: 0.1-0.2)	’10percentage’ ’12percentage’ ’collegeGPA’ ’English’ ’Logical’ ’Quant’ ’Domain’
SelectKbest (With K = 7)	’10percentage’ ’12percentage’ ’collegeGPA’ ’English ’Logical’ ’Quant’ ’Domain’

TABLE II
Feature Selected using Correlation matrix and SelecKbest

By combining the results of both techniques, we were able to select the most relevant features for our predictive model as shown in TABLEII

V. Experiment Set Up and Result

After selecting the features using feature selection techniques, the data was split into training and validation sets using the ’train-test-split’ method. We allocated 20 percent of the data for validation. Later on the split data was experimented with different regression models

- **Linear Regression**: Linear regression is a simple and widely used statistical technique for modeling the relationship between a dependent variable (target) and one or more independent variables (features). In a simple linear regression model, there is only one independent variable, while in a multiple linear regression model, there are multiple independent variables.

- **Decision Tree**: A decision tree is a supervised learning algorithm used for both classification and regression tasks. It works by recursively partitioning the data into subsets based on the values of the features.
- **Random Forest**: Random Forest is an ensemble learning method that encompasses this model includes: Ensemble Learning, Bootstrap Sampling, Random Feature Selection, Decision Tree Construction and Regression(Voting)

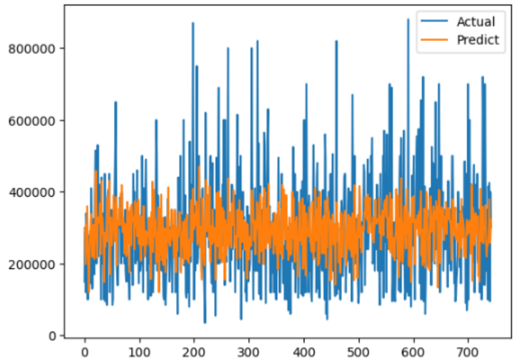

Fig. 1. Result of Linear Regression Model.

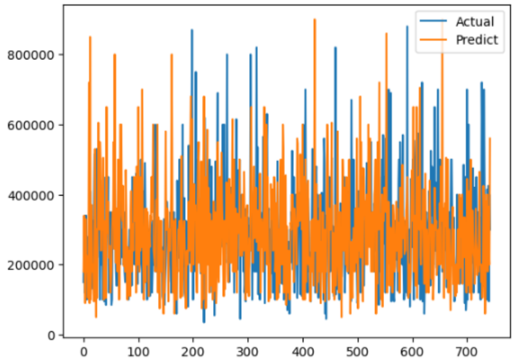

Fig. 2. Result of Decision Forest Model.

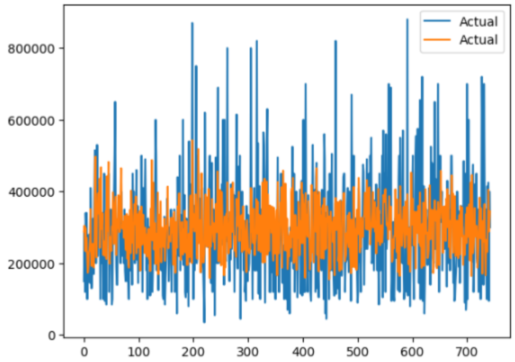

Fig. 3. Result of Decision Forest Model.

The above figures showcases the outcome of the utilized models against the training split of the train data set. The accuracy and precision of the outcomes are discussed in the Result section.

VI. Result Discussion

In order to compare the performance of the models, we evaluated three key indicators: value, Mean Squared Error (MSE), and Mean Absolute Error (MAE).

- R2 value measures the proportion of the variance in the dependent variable that is predictable from the independent variables. A higher R2 value indicates a better fit of the model to the data.
- MSE measures the average squared difference between the actual and predicted values. Lower MSE values indicate better model performance.
- MAE measures the average absolute difference between the actual and predicted values. Lower MAE values indicate better model performance.

Below is a table illustrating the three indicating values of the three separate models.

Model	r2 score	MSE	MAE
Linear Regression	0.208	1.607×10^10	9.645×10^4
Decision Tree	-0.669	3.385×10^10	1.402×10^5
Random Forest	0.169	1.685×10^10	1.003×10^5

TABLE III
Comparison of the different models' result

Among the three models considered, Linear Regression outperformed both Decision Tree and Random Forest in terms of R2 score, Mean Squared Error (MSE), and Mean Absolute Error (MAE). Despite its simplicity, the Linear Regression model achieved a relatively higher R2 score of 0.208 and lower MSE and MAE values compared to the other models.

This indicates that the Linear Regression model provides a better fit to the data and yields more accurate predictions of the annual salary of engineering candidates in our dataset. Where as Decision Tree yielded the least accurate data. As shown in Fig.2, almost all the predicted data points was more than the actual data points resulting in a R2 score of -0.669. While Random Forest models might be more complex and capable of capturing non-linear relationships in the data, the straightforwardness and interpretability of the Linear Regression model make it a suitable choice for this particular prediction task.

Furthermore as outlined in TABLEIII Linear Regression model outperforms both the Decision Tree and Random Forest models in terms of Mean Squared Error (MSE) and Mean Absolute Error (MAE). The Linear Regression model achieved a MSE of 1.607× $10^10 whicharelowerthantheMSEandMAEvaluesoftheDecisionTree$

Linear Regression's ability to provide easily interpretable coefficients for each feature allows us to understand the relationship between the independent variables and the target variable more intuitively. Additionally, Linear Regression is less prone to overfitting, making it more robust when dealing with smaller datasets or datasets with fewer features.

Considering the balance between model performance and interpretability, we conclude that the Linear Regression model is the most suitable choice for predicting the annual salary of engineering candidates in our dataset.

VII. Conclusion

.In this study, we explored the use of machine learning models to predict the annual salary of engineering candidates based on their profile information. Despite our efforts, the performance of the models, were not up to the mark. While Linear Regression provided a simple and interpretable solution, it failed to capture the complex relationships present in the data. Moving forward, more sophisticated modeling techniques and feature engineering methods need to be explored to improve the predictive performance of the model. Additionally, incorporating more relevant features and possibly exploring other machine learning algorithms could lead to better results.Overall, this study serves as a foundation for future research in the field of predicting employment outcomes for engineering candidates in India, but it also highlights the need for further investigation and refinement of the predictive models.