Predictive Modeling of Annual Job Salary for Engineering Candidates in India: A Machine Learning Approach

Sharmin Islam Shroddha
Department of Computer Science and Engineering
Independent University,Bangladesh
Dhaka, Bangladesh
2421194@iub.edu.bd

I. Introduction

The AMEO - Aspiring Minds' Employment Outcomes 2015 dataset provides a unique opportunity to explore the relationship between various candidate profile characteristics and their subsequent employment outcomes in India. The dataset contains a comprehensive set of information about engineering candidates. The contemporary job market demands efficient methods for assessing and predicting employment outcomes.

The goal of this project is to construct machine learning models that can accurately predict the annual job salary of engineering candidates based on their profile characteristics. By leveraging the AMCAT scores and other relevant features, we aim to develop predictive models that can provide insights into the factors that drive salary outcomes in the Indian job market. This study contributes to the existing body of research on employment outcomes and career development by applying machine learning techniques to a large and diverse dataset. The results of this project will provide valuable insights for educators, career counselors, and employers seeking to understand the factors that influence a candidate's career success and salary potential.

II. Dataset Description

The Aspiring Minds' Employment Outcomes (AMEO) 2015 dataset serves as the cornerstone of our analysis. This dataset encapsulates a comprehensive array of information about engineering candidates in India, encompassing both their profile information and employment outcome details. The dataset comprises a total of 3998 data points.

To facilitate model training and validation, we split the dataset into training and validation sets. We utilize 80% of the data for training and the remaining 20% for validation. Additionally, we have a separate test dataset reserved exclusively for evaluating the performance of our trained models. This test dataset consists of 1500 test case.

The dataset includes:

- Scores on Aspiring Minds' AMCAT, a standardized test of job skills, comprising cognitive, domain, and personality assessments.
- Personal information such as gender, date of birth, etc.
- Pre-university information including high school grades and location.
- University information such as GPA, college major, and college reputation proxy.
- Demographic information such as the location of the college and candidates' permanent location.
- Employment outcome information including the first job's annual salary, title, and location.

This dataset stands out due to its unique combination of standardized job test scores and employment outcome data, providing a rich source of information for predictive modeling tasks.

III. Pre-processing Steps

The pre-processing of the dataset involves several crucial steps to ensure its quality and suitability for machine learning model training. The steps undertaken are as follows:

- **Handling Unnecessary Columns:** The 'Unnamed: 0' column is removed from both the training and test datasets as it doesn't contribute to the predictive task. Additionally, columns such as 'DOJ', 'DOL', 'Designation', and 'JobCity' are dropped from the training dataset, as they do not have values in the test dataset and are not relevant for predicting salary.
- **Removing Irrelevant Features:** Columns such as '10board', '12board', 'CollegeID', 'CollegeTier', 'CollegeCityID', 'CollegeCityTier', and 'CollegeState' are deemed irrelevant for predicting salary and thus dropped from both the training and test datasets.
- **Handling Missing Values:** Missing values in the dataset are addressed by identifying and handling them accordingly. Graduation year columns containing 0 are treated as

missing values and imputed with the mode of the column. The 'DOB' column is converted to 'BirthYear', and the age of the candidate at the time of high school graduation ('12GradAge') and college graduation ('GradAge') are calculated and added as new features. Missing values in the 'Domain' column are dropped after attempting to impute them using related columns.

- **Updating GPA Scale:** Updated the college GPA column to ensure consistency by converting GPAs less than 10 to a scale of 100.
- **Feature Encoding:** The 'Gender' column is encoded using LabelEncoder, and one-hot encoding is applied to the 'Degree' and 'Specialization' columns to convert categorical features into numerical representations.
- **Numerical Feature Scaling:** Numerical features are scaled using StandardScaler to bring them to a similar scale, which aids in model convergence and performance.
- **Handling Outliers:** Data points where the salary exceeds 10,00,000 are removed from the training dataset to mitigate the influence of outliers on model training. we filter out rows where the 12th-grade percentage ('12percentage') exceeds 42 and where the college GPA ('collegeGPA') is greater than 53. These thresholds are determined empirically based on the distribution of the data and domain knowledge. By removing these outliers, we ensure that our models are trained on a more representative and reliable subset of the data.

TABLE I
DATA SUMMARY AFTER PRE-PROCESSING

Column	Non-Null Count	Dtype
Salary	3948	int64
10percentage	3948	float64
12percentage	3948	float64
collegeGPA	3948	float64
English	3948	float64
Logical	3948	float64
Quant	3948	float64
Domain	3948	float64
conscientiousness	3948	float64
agreeableness	3948	float64
extraversion	3948	float64
nueroticism	3948	float64
openess_to_experience	3948	float64
BirthYear	3948	int32
12GradAge	3948	float64
GradAge	3948	float64
Gender_encoded	3948	int64
Degree_M.Sc. (Tech.)	3948	bool
Degree_M.Tech./M.E.	3948	bool
Degree_MCA	3948	bool
Specialization_CS	3948	bool
Specialization_EC	3948	bool
Specialization_EL	3948	bool
Specialization_ME	3948	bool
Specialization_other	3948	bool

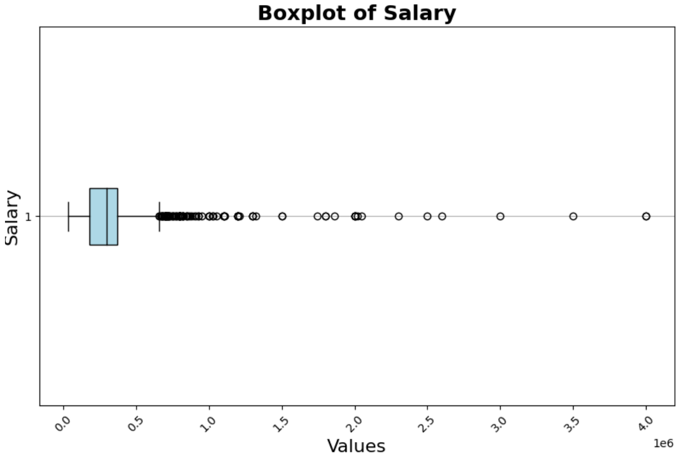


Fig. 1. Boxplot of Salary

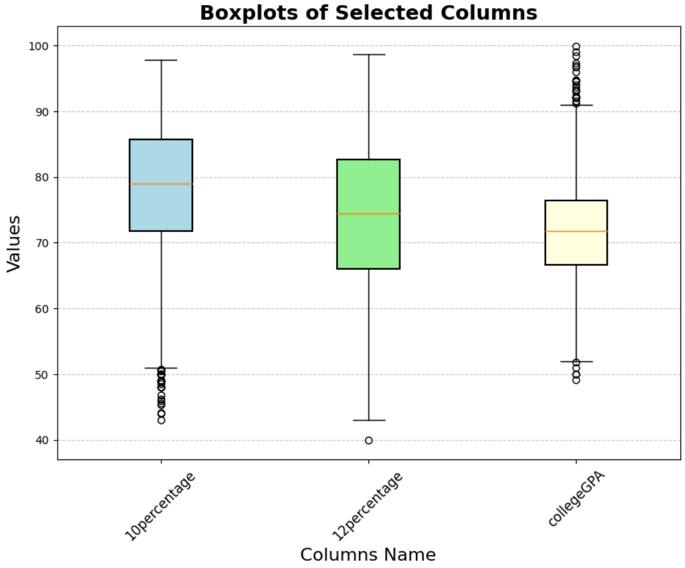


Fig. 2. Boxplot of Selected Columns

IV. Model Description And Feature Selection

In this section, we describe the machine learning models employed for predicting annual job salaries and the feature selection process undertaken to identify the most relevant predictors.

A. *Model Description:*

We explore the following machine learning algorithms for regression tasks:

- **Linear Regression:** A fundamental supervised learning algorithm that models the relationship between the independent variables and the target variable by fitting a linear equation.

- **Decision Tree Regressor:** A non-parametric supervised learning algorithm that recursively partitions the data into subsets based on the features, creating a tree-like structure to make predictions.
- **Random Forest Regressor:** An ensemble learning method that combines multiple Decision Trees to improve predictive performance by training each tree on a random subset of the data and features.
- **XGBoost Regressor:** A powerful gradient boosting algorithm that sequentially builds a series of decision trees, where each tree corrects the errors of the previous ones, known for its efficiency and performance.

These models were selected to capture different approaches to regression, from the simple linear model to the more complex ensemble methods. By comparing the performance of these models on the dataset, the most suitable model(s) for predicting the First job annual salary can be identified.

B. *Feature Selection:*

Explored various feature selection techniques to identify the most important predictors of annual job salaries. These techniques include:

- **Correlation Analysis:** Computed the correlation matrix to assess the linear relationship between each feature and the target variable, annual job salary. Features with high absolute correlation values are considered as potential predictors.
- **SelectKBest with ANOVA F-value:** Utilized the SelectKBest method with ANOVA F-value scoring to select the top k features that exhibit the strongest relationship with the target variable. This method evaluates the individual predictive power of each feature.

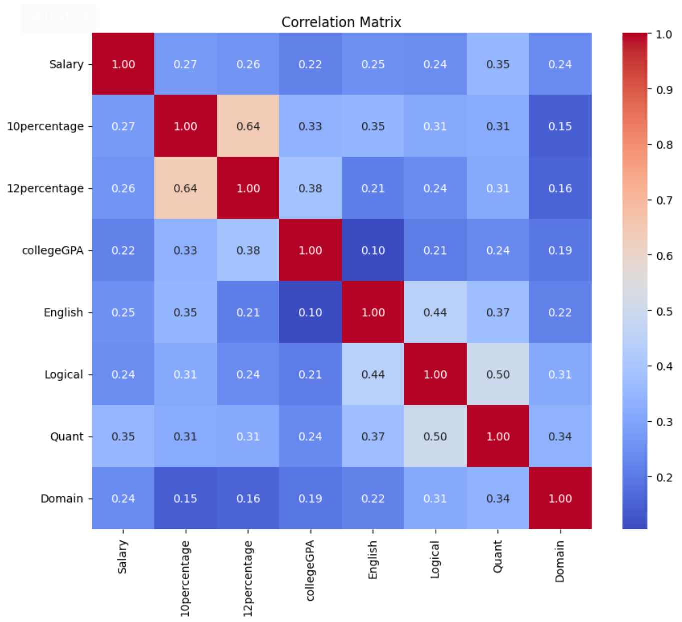
Based on the results obtained from the feature selection techniques, we identify two sets of features:

- **Strong Correlation Features:** These features exhibit strong correlation with the target variable and include 'Quant', '10percentage', '12percentage', 'English', 'Logical', 'Domain', and 'collegeGPA'.
- **Moderate Correlation Features:** These features demonstrate moderate correlation with the target variable and include 'BirthYear', 'Degree_MCA', 'conscientiousness', 'agreeableness', 'nueroticism', 'GradAge', and 'Gender_encoded'.

Trained the models using the strong correlation features and with the union of strong and moderate correlation features.

V. Experiment Setup and Results

In this section, we outline the setup of our experiments, including the training and evaluation of machine learning models, and present the results obtained.



	Salary	10percentage	12percentage	collegeGPA	English	Logical	Quant	Domain
Salary	1.00	0.27	0.26	0.22	0.25	0.24	0.35	0.24
10percentage	0.27	1.00	0.64	0.33	0.35	0.31	0.31	0.15
12percentage	0.26	0.64	1.00	0.38	0.21	0.24	0.31	0.16
collegeGPA	0.22	0.33	0.38	1.00	0.10	0.21	0.24	0.19
English	0.25	0.35	0.21	0.10	1.00	0.44	0.37	0.22
Logical	0.24	0.31	0.24	0.21	0.44	1.00	0.50	0.31
Quant	0.35	0.31	0.31	0.24	0.37	0.50	1.00	0.34
Domain	0.24	0.15	0.16	0.19	0.22	0.31	0.34	1.00

Fig. 3. A Subset of the Correlation Matrix

TABLE II
Top Features Based on ANOVA F-value and Correlation with Salary

Feature	ANOVA F-value Score	Correlation with Salary
Quant	551.93	0.3503
10percentage	316.61	0.2725
12percentage	287.40	0.2606
English	261.32	0.2492
Logical	245.64	0.2421
Domain	236.16	0.2376
collegeGPA	199.57	0.2194
BirthYear	112.35	0.1664
Degree_MCA	20.62	0.0721
conscientiousness	17.27	0.0660
agreeableness	14.37	0.0602
nueroticism	13.02	0.0574
Gender_encoded	6.75	0.0413
GradAge	6.25	0.0398
Specialization_EC	3.99	0.0318
extraversion	2.66	0.0259
Specialization_EL	2.43	0.0248
Specialization_CS	2.33	0.0243
Degree_M.Tech./M.E.	1.72	0.0209
Specialization_other	1.54	0.0197
openess_to_experience	0.66	0.0129
12GradAge	0.19	0.0069
Degree_M.Sc. (Tech.)	0.06	0.0040
Specialization_ME	0.01	0.0013

A. *Experiment Setup:*

- **Data Splitting:** We split the dataset into training and validation sets using a predefined percentage split. Specifically, we allocate 80% of the data to the training set and 20% to the validation set. The training set is used to train the models, while the validation set is used to assess their performance.
- **Feature Selection:** Two sets of features were used for

model training: strong correlation features and a combination of strong and moderate correlation features. These feature sets were selected based on their correlation with the target variable.
- **Model Training:** Each of the selected machine learning models is trained using the training data. Hyperparameters are tuned using techniques such as cross-validation to optimize model performance.
- **Evaluation Metrics:** To evaluate the performance of the models, we employ metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R$\hat{2}$). These metrics provide insights into the accuracy and generalization capability of the models.

B. Results:

We present the performance of each model on the validation set using the evaluation metrics mentioned above. This allows for comparison and identification of the best-performing model. Based on the provided cross-validation RMSE scores(Table III), the model with the lowest mean RMSE is the Linear Regression. A lower RMSE indicates better performance, so the Linear Regression has the best performance among the models tested. Residual plots(Fig. 5) were generated for each model to visualize the distribution of residuals and assess the model's performance. Scatter plots of actual versus predicted values(Fig. 4) were created for each model to visually inspect the model's predictive accuracy.

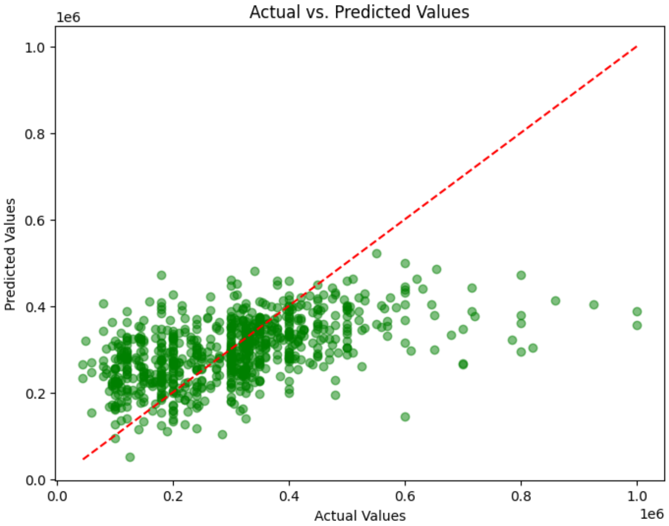


Fig. 4. Actual Value Vs Predicted Value Plotting

VI. Discussion on the Results

In this section, we discuss the results obtained from the experiments and provide insights into the performance of the machine learning models in predicting annual job salaries.

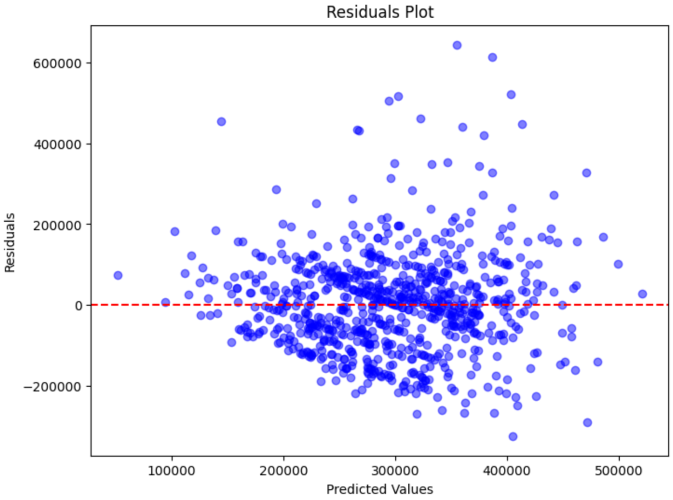


Fig. 5. Residual Plotting

A. Model Performance

Linear Regression Dominance: The Linear Regression model consistently outperformed the other models across all evaluation metrics, indicating its superior predictive capability. This suggests that a linear relationship between input features and annual job salaries may exist in the dataset.

Impact of Feature Selection: The performance of the models varied based on the set of features used for training. Models trained on a combination of strong and moderate correlation features generally performed better than those trained solely on strong correlation features. This highlights the importance of including a diverse set of features for accurate salary prediction.

B. Interpretation of Metrics

Mean Squared Error (MSE): Linear Regression achieved the lowest MSE, indicating minimal prediction errors compared to the other models. DecisionTreeRegressor exhibited the highest MSE, suggesting poor performance and large prediction errors.

Mean Absolute Error (MAE): Similar to MSE, Linear Regression attained the lowest MAE, signifying smaller absolute prediction errors. DecisionTreeRegressor recorded the highest MAE, indicating larger absolute errors in predictions.

R-squared (R^2) Score: Linear Regression demonstrated the highest R-squared score, indicating better fit and higher explanatory power compared to other models. DecisionTreeRegressor exhibited a negative R-squared score, indicating poor model fit and worse performance than a simple mean prediction.

VII. Conclusion

In conclusion, the experiments demonstrate the effectiveness of machine learning models in predicting annual job salaries. The results underscore the importance of feature selection,

TABLE III
Cross-Validation RMSE Scores and Mean RMSE

Model	Cross-Validation RMSE Scores	Mean RMSE
Linear Regression	[124269.60, 126368.21, 123672.22, 129861.34, 133494.63]	127533.20
Decision Tree Regressor	[184652.86, 182493.28, 185553.61, 184396.37, 192598.90]	185939.00
Random Forest Regressor	[126781.50, 129491.59, 127662.36, 133940.89, 137588.39]	131092.95
XGB Regressor	[132443.67, 135238.55, 137262.71, 140977.87, 142886.49]	137761.86

TABLE IV
Model Evaluation Metrics - Using Strong Correlation Features

Model	Mean Squared Error (MSE)	Mean Absolute Error (MAE)	R-squared (R2) Score
XGBRegressor	21053634603.92	109964.32	0.0068
LinearRegression	17346867018.34	98660.66	0.1817
DecisionTreeRegressor	38909374683.54	148779.75	-0.8355
RandomForestRegressor	18407687287.97	101350.19	0.1316

TABLE V
Model Evaluation Metrics - Using Strong And Moderate Correlation Features

Model	Mean Squared Error (MSE)	Mean Absolute Error (MAE)	R-squared (R2) Score
Linear Regression	15568039608.92	92666.99	0.2656
XGBRegressor	19338971199.82	103926.55	0.0877
DecisionTreeRegressor	35280640506.33	138475.95	-0.6643
RandomForestRegressor	16853553892.41	96984.43	0.2050

model simplicity, and interpretability in achieving accurate predictions. The insights gained from this study can inform decision-making processes in career planning and salary negotiation.