Assignment 1

Regression Models

Shashikanth Senthil Kumar
Student ID: 25218722

2024 SPRING

36106 - Machine Learning Algorithms and Applications
Master of Data Science and Innovation
University of Technology of Sydney

Table of Contents

1. Business Understanding

a. Business Use Cases

The project involves predicting the future salary of engineering graduates from various colleges. The primary business use cases include:

Career Counseling and Planning: Educational institutions can use salary predictions to provide career advice to students. This helps students make informed decisions about their career paths and academic specializations.

Talent Acquisition and Management: Companies can use the salary predictions to benchmark their compensation packages and attract top talent. Understanding expected salaries helps in structuring competitive offers and planning salary budgets.

Educational Institution Evaluation: Colleges and universities can assess how well their programs prepare students for high-paying roles, providing insights into the effectiveness of their educational offerings.

Challenges:

- **Feature Engineering:** Predicting salaries involves multiple factors, such as academic performance, college reputation, and individual skills. Selecting and creating relevant features from these factors is a complex task.
- **Model Generalization:** Ensuring the model performs well on new data.

Opportunities:

- **Data-Driven Decisions:** Provides objective insights for better decision-making.
- **Improved Competitiveness:** Aligns compensation and program offerings with market standards.
- **Personalized Recommendations:** Offers tailored insights based on individual profiles.

b. Key Objectives

Develop Accurate Salary Predictions: Train regression models to predict the salaries of engineering graduates based on various features such as academic performance, college tier, and skills.

Evaluate Model Performance: Assess the accuracy of different regression models using metrics like RMSE (Root Mean Squared Error) to ensure reliable predictions.

Provide Actionable Insights: Generate insights from the model results to support decision-making processes in talent acquisition, career counseling, and educational evaluation.

Stakeholders and Their Requirements

1. **Career Counselors and Educational Advisors**: Require insights into future salaries to guide students in making informed choices about their education and career paths.
2. **Human Resources and Talent Acquisition Teams**: Need accurate salary predictions to design competitive compensation packages and attract high-caliber candidates.
3. **Educational Institutions**: Seek to evaluate the effectiveness of their programs and improve their offerings based on the predicted salary outcomes of their graduates.

Project Goals

- **Model Training**: Train multiple regression models (Multivariate Linear Regression, Ridge Regression, Lasso Regression) to predict salaries accurately.
- **Performance Assessment**: Use RMSE and other metrics to evaluate the performance of these models and select the best-performing one.
- **Business Impact**: Ensure that the predictions and insights lead to better decision-making, enhanced strategic planning, and improved alignment between education and employment outcomes.
- **Recommendations**: Based on the model outcomes, provide recommendations for improving educational programs, compensation strategies, and career counseling practices.

2. Data Understanding

a. Dataset Overview

The dataset used for this project is aimed at predicting the salary of engineering students based on a variety of academic, demographic, and personality-related features. The dataset is split into three distinct parts:

- **Training dataset:** Contains 1803 entries used to train the model.
- **Validation dataset:** Consists of 901 entries for tuning hyperparameters and validating model performance.
- **Testing dataset:** Contains 902 entries to evaluate the final model's performance.

Data Collection and Sources

The dataset appears to be collected from multiple sources related to students' academic records, personality traits, and demographic details. However, the specific sources of the data or the methods used for data collection are not provided in the Brief. This lack of information on data provenance might affect the interpretability and reliability of the model's predictions.

b. Variables/Features and Their Relevance

Key features present in the dataset include:

1. **CollegeGPA:** A significant predictor of salary, as academic performance often correlates with job prospects.
2. **12percentage & 10percentage:** These features capture past academic performance, which may impact the salary.
3. **Test Scores (English, Logical, Quantitative, ComputerProgramming):** Represent analytical and reasoning abilities that could influence employability and salary.
4. **Personality Traits:** Scores related to various traits, such as teamwork and problem-solving, may also impact salary predictions.
5. **Specialization:** The student's area of focus in engineering, which could directly affect salary prospects depending on the demand for specific specializations in the job market.

c. Exploratory Data Analysis (EDA) Insights

- **Salary Distribution:**
    - **Training Dataset:** The salary range is narrow, with minimal skewness, suggesting similar salaries for students in this dataset (see Fig-1).
    - **Validation Dataset:** A broader salary range with a slight skew toward higher salaries (see Fig-2).
    - **Testing Dataset:** The testing dataset shows a wide salary range, with significant right skew due to high outliers (see Fig-3).
- The differences in salary distribution among datasets raise concerns about the model's performance on unseen data, particularly for outlier cases in the testing dataset.
- **Academic Performance:**
    - **College GPA:** Most students have a GPA between 65 and 80, with a concentration around 70-75.
    - **12percentage:** The average percentage score varies across datasets (e.g., Training: 74.79%, Validation: 71.25%, Testing: 76.08%). This variability could introduce challenges in predicting salaries.
- **Personality Traits & Test Scores:**
    - **Quant:** The average quantitative score varies across datasets (Training: 518.48, Validation: 467.18, Testing: 548.21). This variability, along with the presence of extreme values, could impact the model's ability to predict salaries accurately.

d. Data Limitations

Several limitations were identified during the data exploration process:

- **Irrelevant Features:** The dataset contains irrelevant personal identifiers such as passport numbers, DOB, and names, which need to be removed to avoid privacy concerns and prevent data leakage.
- **Data Discrepancy:** There are significant variations in salary distributions across training, validation, and testing datasets, which could impact the model’s generalizability.
- **Potential Outliers:** In the testing dataset, extreme salary values could skew predictions, requiring outlier management.

Despite these limitations, the dataset offers valuable features related to academic performance, test scores, and personal characteristics. Addressing these challenges through proper data preprocessing, feature engineering, and outlier management will be crucial for enhancing model performance and ensuring robust predictions.

3. Data Preparation

The data preparation process ensures that the dataset is clean, relevant, and appropriately transformed, leading to more effective and accurate model predictions.

a. Data Cleaning

i. Dataset Copying:

- **Step Taken:** Datasets were copied to ensure that the original data remained unchanged throughout the cleaning process.
- **Purpose:** Preserves the integrity of the original data and allows for safe experimentation during cleaning.

ii. Handling Missing Values:

- **Step Taken:** The datasets were checked for missing values, but none were found.
- **Purpose:** Ensures data integrity and avoids issues that missing values could cause during modeling.

iii. Fixing Data Inconsistencies:

- **Step Taken:** Categorical data entries were standardized to ensure uniformity across all datasets and Inconsistent values such as -1 were replaced with 0 .
- **Purpose:** Standardizing text and correcting inconsistencies ensures accurate and reliable data for analysis and modeling.

iv. Removing Duplicated Values:

- **Step Taken:** Checked for and confirmed that there were no duplicated values in any dataset.
- **Purpose:** Ensures that each data point is unique and prevents bias in the analysis.

b. Feature Selection

i. Domain Knowledge Selection:

- **Step Taken:** Removed irrelevant, redundant, or sensitive features such as personal identifiers (e.g., passport_number, phone_number) and contact details from the datasets.
- **Purpose:** Eliminates features that do not contribute to salary prediction and adhere to privacy and ethical standards.

ii. Correlation Analysis:

- **Step Taken:** Performed correlation analysis (see Fig-4) to identify and retain features with strong positive correlations with the target variable, Salary. Features with weak or negative correlations were excluded.
- **Purpose:** Focuses on features that are most likely to influence salary outcomes, enhancing model performance and interpretability.

c. Feature Engineering

i. Interaction Terms:

- **Step Taken:** Created new features to capture interactions between existing features:
    - GPA_Quant_Interaction: Interaction between college GPA and Quantitative skills.
    - 10_12_Percentage_Interaction: Interaction between 10th-grade percentage and 12th-grade percentage.
- **Purpose:** Captures compounded effects of multiple features, potentially improving model accuracy by revealing hidden relationships.

ii. Categorical Transformation:

- **Step Taken:** Categorized college GPA into Low, Medium, or High categories.
- **Purpose:** Simplifies the relationship between GPA and salary, potentially uncovering patterns not evident with raw numerical values.

d. Data Preprocessing

i. Dataset Splitting:

- **Step Taken:** Split the datasets into training, validation, and test sets. Separated the target variable, Salary, from the feature set.
- **Purpose:** Allows for proper training, validation, and testing of the model to ensure it generalizes well to unseen data.

ii. Label Encoding of Categorical Features:

- **Step Taken:** Applied label encoding to convert categorical features into numerical values.
- **Purpose:** Ensures that categorical data can be effectively used in machine learning algorithms that require numerical input.

iii. Standardization of Features:

- **Step Taken:** Standardized features to have a mean of 0 and a standard deviation of 1.
- **Purpose:** Ensures that all features contribute equally to the model, improving performance and convergence of algorithms sensitive to feature scaling.

iv. Conversion to DataFrame Post-Scaling:

- **Step Taken:** Converted scaled data arrays back to DataFrames with original column names.
- **Purpose:** Enhances readability, facilitates integration with tools and libraries, and preserves metadata for subsequent processing.

v. Handling Outliers and Imbalanced Data:

- **Outliers:** Outlier detection techniques like IQR and z-score were attempted, but the removal of outliers led to significant data loss, negatively impacting model performance. As a result, no explicit outlier handling was implemented.
- **Imbalanced Data:** No specific handling of imbalanced data was described. Techniques such as resampling or using algorithms robust to class imbalance might be applied if necessary.

4. Modeling

a. Machine Learning Algorithms Used

- **Multivariate Linear Regression**: Predicts a continuous outcome (salary) based on multiple features with a linear relationship. It serves as a straightforward, interpretable baseline model.
- **Ridge Regression**: Enhances linear regression by adding L2 regularization, which reduces the impact of less important features and helps prevent overfitting, especially useful with many predictors.
- **Lasso Regression**: Utilizes L1 regularization to perform feature selection by shrinking some feature coefficients to zero, simplifying the model and focusing on the most relevant features.

b. Rationale Behind Algorithm Selection

- **Multivariate Linear Regression**: Provides a clear and simple model to understand relationships between features and salary.
- **Ridge Regression**: Addresses issues like multicollinearity and overfitting, improving model robustness with regularization.
- **Lasso Regression**: Combines feature selection with regularization, helping to build a more interpretable and efficient model by excluding less important features.

c. Hyperparameter Tuning

- **Linear Regression:** The tuning focused on the fit_intercept parameter, with values set to [True, False]. Instead of using grid search, a for loop was employed to test these values systematically. This approach provided insights into how the inclusion or exclusion of an intercept term affected model performance across training, validation, and test datasets.
- **Ridge and Lasso Regression:** Although grid search was initially planned to optimize the alpha parameter, which controls regularization strength, a for loop was used to evaluate different values of alpha [0.01, 0.1, 1, 10, 100]. Opting for a direct evaluation using loops for hyperparameter tuning provided a detailed understanding of how each parameter setting impacted model performance.

d. Model Selection Criteria

- **Evaluation Metric**: The performance of each model was evaluated using RMSE, which is a key metric for assessing the accuracy of the model's predictions. Lower RMSE values indicate better model performance.
- **Model Performance**: Models were assessed based on their RMSE across training, validation, and test datasets. This evaluation provided insights into how well each model generalizes to unseen data and helped in selecting the best-performing model.
- **Final Selection**: The final model was chosen based on the comparison of RMSE values for different hyperparameter settings. The model with the lowest RMSE on the test set was considered the best, as it demonstrated the best generalization capability.

5. Evaluation

a. Results and Analysis

Experiment 1: Impact of Intercept Term in Linear Regression

- **Hypothesis**: Including the intercept term (fit_intercept=True) improves model performance.
- **Results**:
  - **fit_intercept=True**: Achieved significantly lower RMSE values across all datasets (Training: 2225.97, Validation: 5489.77, Test: 15071.06).
  - **fit_intercept=False**: Demonstrated high RMSE values, indicating poor performance (Training: 85744.23, Validation: 82588.62, Test: 94019.67).
- **Analysis**:
  - The model with fit_intercept=True (see Fig-5) achieved much lower RMSE values across all datasets (training, validation, and test) compared to fit_intercept=False. This demonstrates that the intercept term is vital in linear regression models, as it significantly improves the model's accuracy by adjusting the predictions properly to fit the data.

Experiment 2: Impact of Alpha in Ridge Regression

Hypothesis: Higher alpha values will enhance predictive performance by reducing overfitting.

Results:

- **alpha=0.01**: Relatively high RMSE (Training: 2225.97, Validation: 5489.76, Test: 15071.05).
- **alpha=0.1**: Slightly better performance (Training: 2225.97, Validation: 5489.71, Test: 15070.94).
- **alpha=1**: Improved performance (Training: 2225.99, Validation: 5489.22, Test: 15069.96).
- **alpha=10**: Best balance between training and validation RMSE (Training: 2226.72, Validation: 5487.80, Test: 15065.85).
- **alpha=100**: Increased RMSE due to underfitting (Training: 2228.57, Validation: 5496.46, Test: 15070.79).

Analysis:

- The results show that increasing alpha values improved model performance up to a point. Specifically, alpha=10 (see Fig-6) provided the best balance between training and validation RMSE, indicating effective regularization and better generalization. However, very high alpha values led to underfitting, as seen with alpha=100, where the RMSE increased due to excessive regularization.

Experiment 3: Impact of Alpha in Lasso Regression

Hypothesis: Higher alpha values will enhance predictive performance by reducing overfitting and improve model generalization.

Results:

- **alpha=0.01**: Relatively high RMSE(Training RMSE: 2225.97, Validation RMSE: 5489.76, Test RMSE: 15071.05).
- **alpha=0.1**: Slightly better performance (Training RMSE: 2225.97, Validation RMSE: 5489.71, Test RMSE: 15070.94).
- **alpha=1**: Improved performance (Training RMSE: 2226.25, Validation RMSE: 5488.43, Test RMSE: 15068.52).
- **alpha=10**: Best balance between training and validation RMSE (Training RMSE: 2228.28, Validation RMSE: 5494.41, Test RMSE: 15072.75).
- **alpha=100**: Increased RMSE due to underfitting (Training RMSE: 2256.16, Validation RMSE: 5578.75, Test RMSE: 15182.69).

Analysis:

- The Lasso Regression model achieves its best performance with alpha=1 (see Fig-7), providing lower RMSE compared to other alpha values. Higher alpha values (e.g., alpha=10 and alpha=100) lead to underfitting and higher RMSE, indicating poor model performance.

Comparison:

- **Linear Regression (fit_intercept=True)**: Provides a baseline model with decent performance but does not include regularization, which might be crucial for managing overfitting.
- **Ridge Regression**: Performs best with moderate alpha values, particularly **alpha=10**, showing effective regularization and generalization.

- **Lasso Regression**: Also performs best with **alpha=1**, offering a balance between regularization and model accuracy. It performs well in feature selection by shrinking some coefficients to zero.

Best Model:

Ridge Regression with alpha=10 is the best model among the three. It achieves the lowest RMSE values on the validation and test datasets compared to Lasso Regression and Linear Regression. Ridge Regression effectively balances bias and variance with moderate regularization, providing reliable predictions while managing overfitting.

Reason for Selection:

- Ridge Regression shows the best balance between model complexity and generalization.
- It performs better in terms of RMSE on the test set compared to the other models.
- Provides reliable predictions and effective regularization, making it the most suitable model for this dataset and problem.

In conclusion, Ridge Regression with alpha=10 is recommended for its superior performance in predicting salary, improving model accuracy, and ensuring effective generalization.

b. Business Impact and Benefits

- **Positive Impacts**: Quant, 10percentage and 12percentage, English, collegeGPA, domain, Computerprogramming, civileng and telecomeng, extraversion, and agreeableness contribute to higher salary predictions. This suggests businesses should prioritize candidates with strong academic, technical, and interpersonal skills.
- **Negative Impacts**: Lower interaction between GPA and quantitative skills, fields like computerscience, conscientiousness, and specific engineering disciplines (e.g., electrical, mechanical), as well as traits like neuroticism and logical reasoning, are associated with lower salary predictions. Employers should be cautious when assessing these features in candidates.
- **Business Use Case**: Companies should focus on hiring candidates with strong quantitative, programming, and communication abilities (see Fig-8). Educational institutions can improve employability by focusing on these key areas.

c. Data Privacy and Ethical Concerns

Data Privacy Implications:

- **Sensitive Information**: The dataset contains personal information such as names, dates of birth (DOB), gender, passport numbers, credit card details, phone numbers, and email addresses. Such sensitive data must be securely handled to protect individual privacy.

Steps Taken for Data Privacy:

- **Sensitive Data Removal**: During the feature selection process, sensitive information, including credit card details, passport numbers, phone numbers, and email addresses, was dropped to ensure no personal data is used in model training.
- **Anonymization**: All identifiable personal information (e.g., first name, last name, and address) was excluded from the analysis to maintain data anonymity and privacy.

Ethical Considerations:

- **Responsible Data Use**: We ensured that data collection, usage, and model deployment adhered to ethical standards, avoiding any form of discrimination or misuse.
- **Bias Mitigation**: To prevent biased predictions, sensitive features such as gender were removed from the dataset during the preprocessing phase.

Potential Negative Impacts:

- **Indigenous Communities**: Care was taken to ensure that the model does not perpetuate biases or inequalities, especially toward marginalized groups, such as Indigenous communities.

6. Conclusion

In this project, we successfully developed a predictive model to estimate the future salaries of engineering graduates using key features such as academic performance, demographics, and personal attributes. By applying Multivariate Linear Regression, Ridge Regression, and Lasso Regression techniques, we evaluated how these factors relate to salary outcomes and assessed the models using the RMSE metric to measure prediction accuracy.

The key findings from our experiments include:

- Ridge Regression with an alpha value of 10 proved to be the best-performing model, offering the most accurate salary predictions with effective regularization.
- Including an intercept term significantly boosted the models' performance, improving their ability to predict salary outcomes.
- Moderate regularization, particularly in Ridge Regression, effectively handled overfitting while maintaining accuracy, whereas high regularization led to underfitting.
- Lasso Regression, although useful for feature selection, did not outperform Ridge Regression in terms of prediction accuracy.

The project met its objective of building a reliable model for salary prediction, and the results can be valuable to stakeholders in career counseling, talent acquisition, and educational evaluation. Institutions can refine their programs to increase graduates' employability, while companies can structure compensation strategies based on expected salary outcomes.

Future Work and Recommendations:

- To further improve the model, future steps could involve experimenting with additional features such as internship experience or industry-specific attributes.
- We also recommend testing more advanced techniques like ensemble methods (e.g., Random Forest, Gradient Boosting) to enhance predictive power.
- Additionally, deploying the model in real-world applications could provide further validation and refinements based on user feedback and new data sources.

7. References

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32. https://link.springer.com/article/10.1023/A:1010933404324

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: With Applications in R*. Springer. https://www.statlearning.com/

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer. https://www.springer.com/gp/book/9781461468486

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267-288. https://onlinelibrary.wiley.com/doi/10.1111/j.2517-6161.1996.tb02080.x

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(2), 301-320. https://onlinelibrary.wiley.com/doi/10.1111/j.1467-9868.2005.00503.x

Friedman, J., Hastie, T., & Tibshirani, R. (2010). *Regularization Paths for Generalized Linear Models via Coordinate Descent*. Journal of Statistical Software, 33*(1), 1-22. https://www.jstatsoft.org/article/view/v033i01

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Blondel, M. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825-2830. http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

8. Appendices

Fig- 1 (Salary Distribution in Training Dataset)

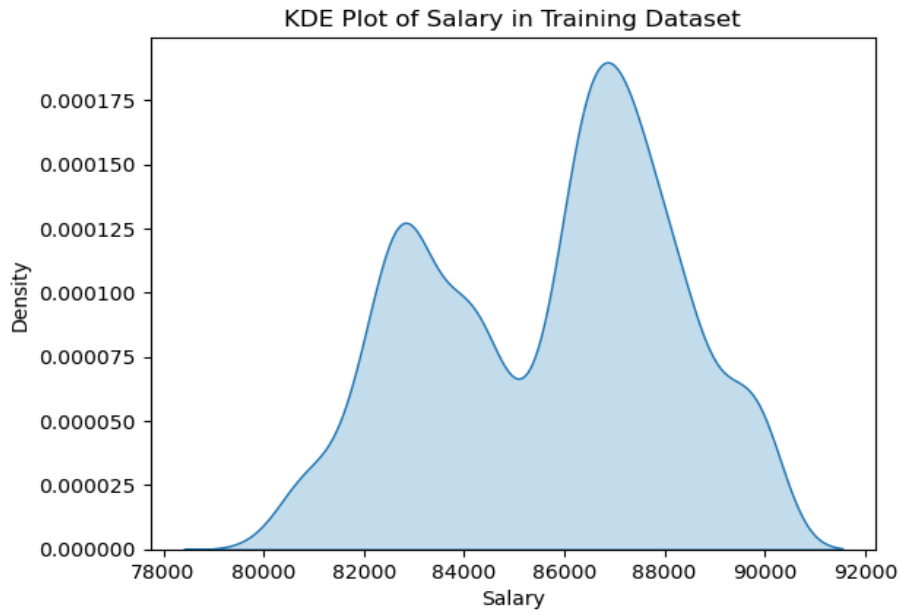

Fig- 2 (Salary Distribution in Validation Dataset)

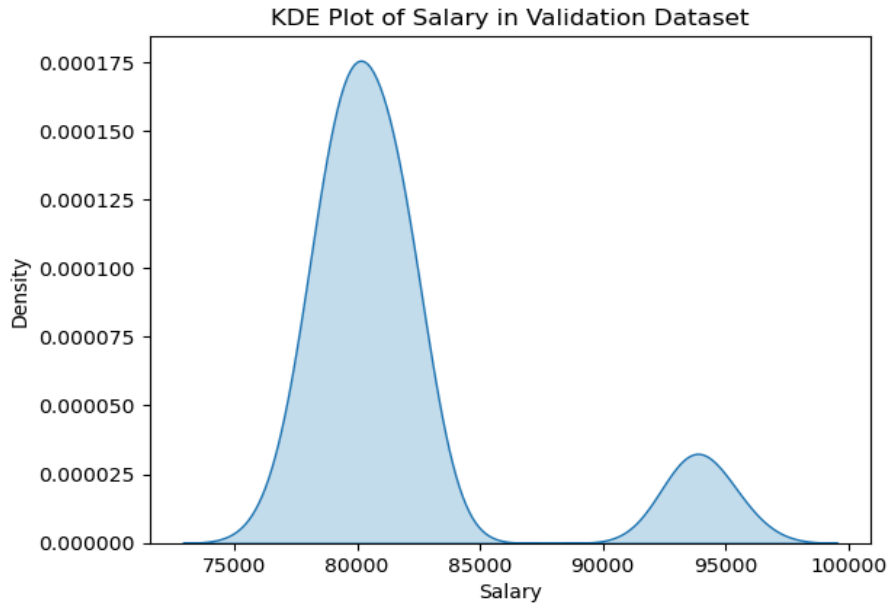

Fig-3 (Salary Distribution in Testing Dataset)

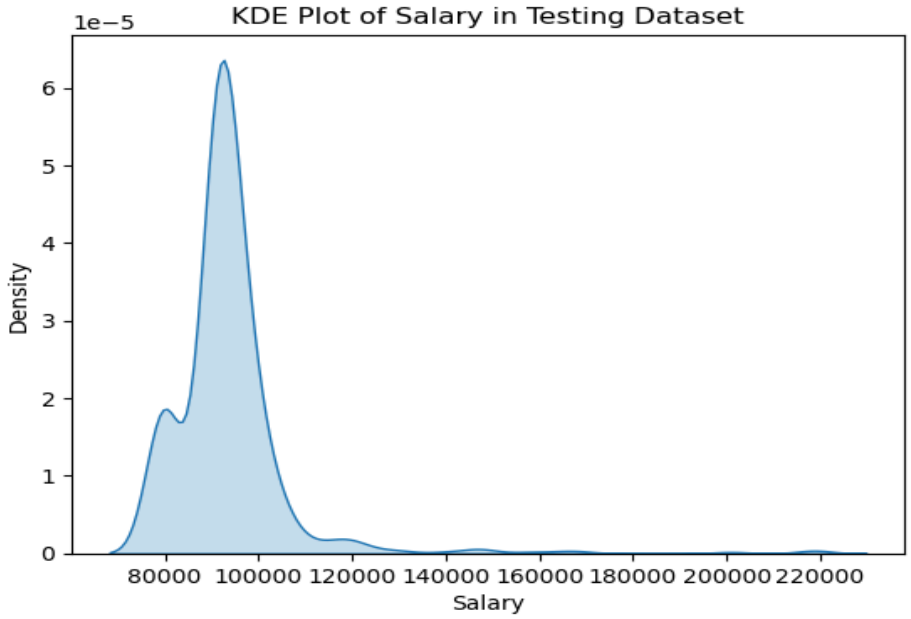

Fig-4 (Correlation Analysis)

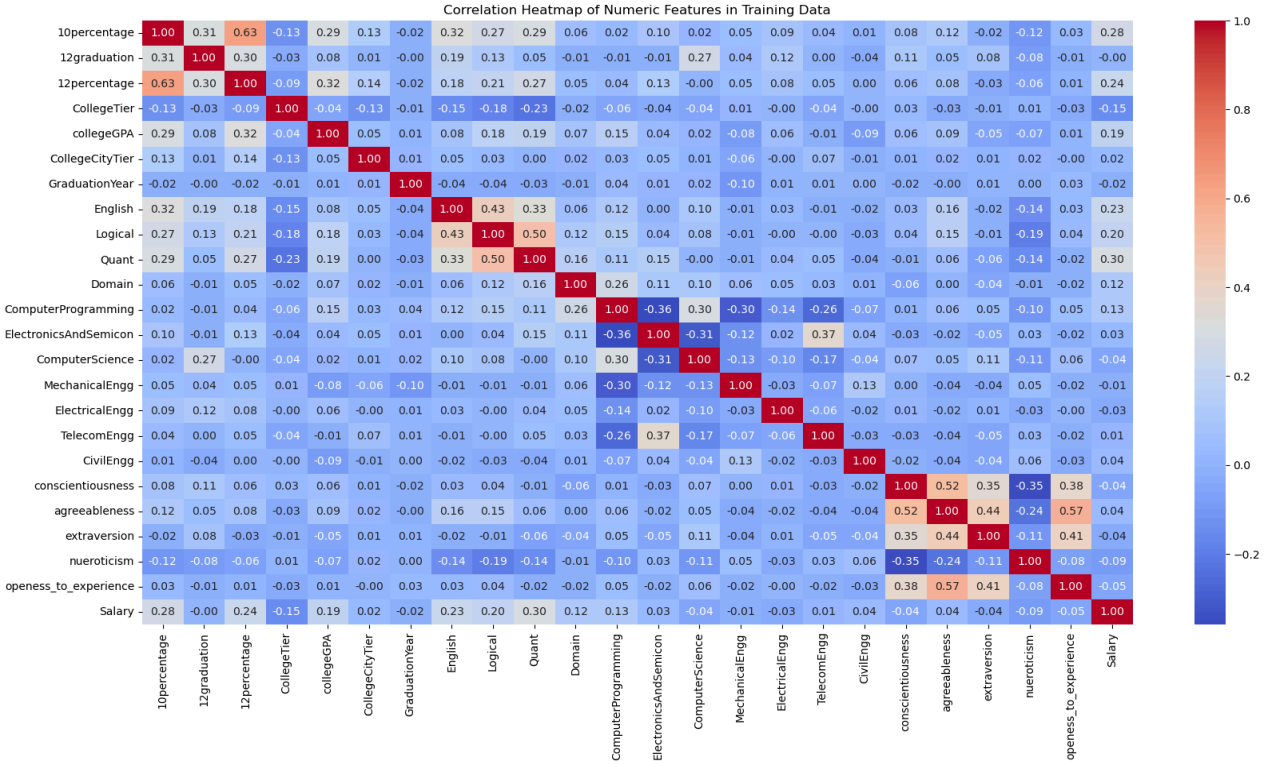

Fig-5(Best Linear Regression model results)

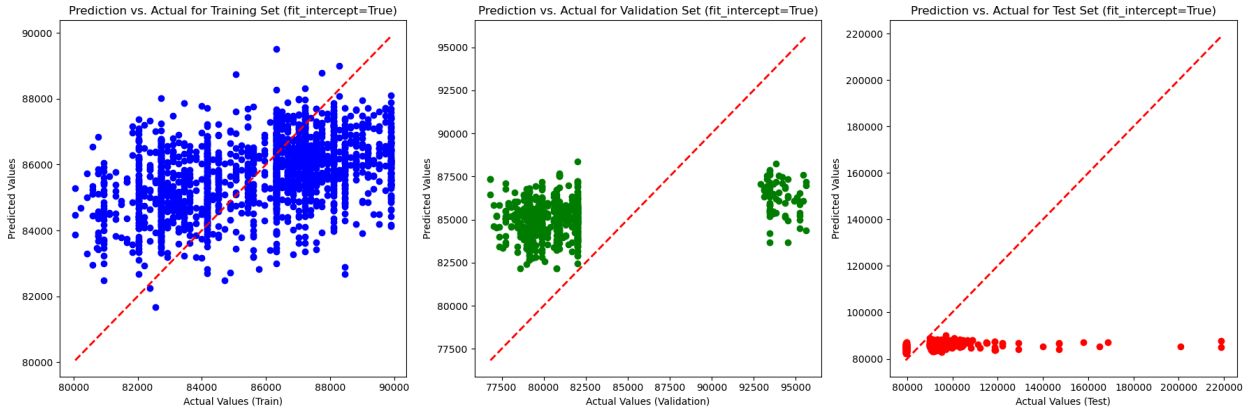

Fig-6(Best RidgeRegression model results)

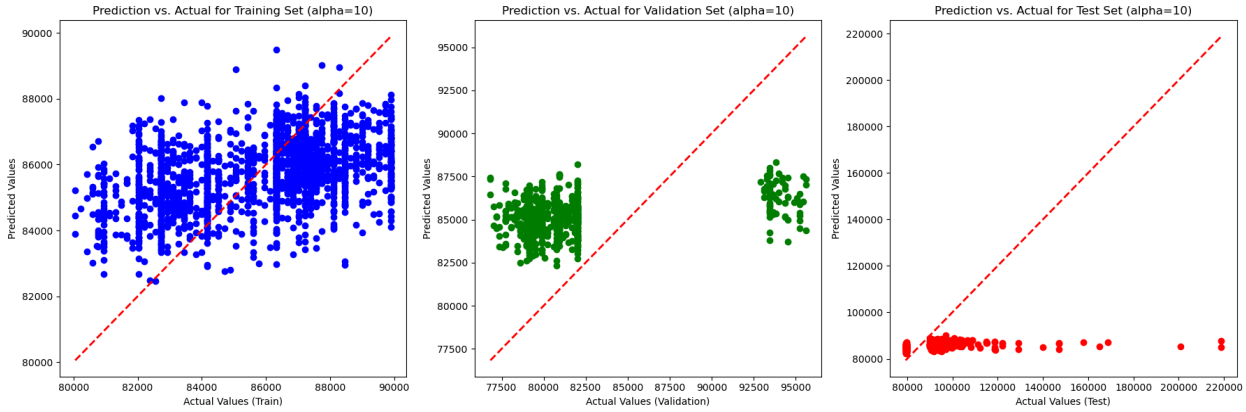

Fig-7 (Best Lasso Regression model results)

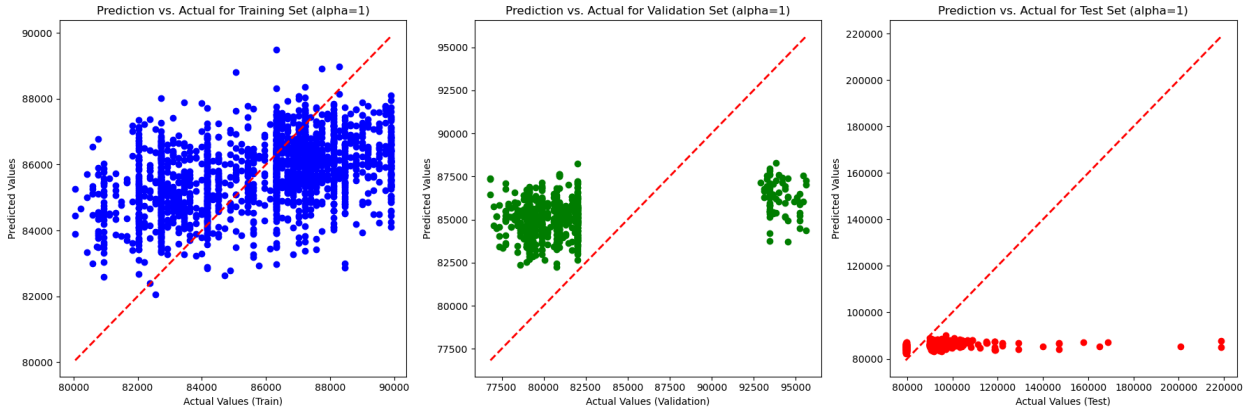

Fig-8 (Impact of features in Ridge Model)