# Department of Statistics

# University of Colombo

## ST4035 - Data science

## Final Report


***Group 05***

**S14304 - M.L.S.R. Abhayawardhana**

**s14379 - J.M.D.Madushan Jaya Sri**

**s14357 - W.A.M.Nuwanjalee**

**s14487 – Agrani Nimshani**

## Problem Statement

Factory workers are the main workforce contributing to an industry's day to day activities. As a result of their efficacy industries are gaining huge profits, benefits as well as losses. Workers can be laborers, team leaders, shift managers, production directors etc. Typically, the workers of a factory operate in separate teams and they work in different shifts. So, from the influence of the supervisor to the mindset of the worker, a lot of factors affect the efficacy of a worker. Therefore, it is important to identify the factors affecting a worker's daily level of efficacy in order to improve their efficacy as well as the revenue of the company.

## Data

This is a synthetic dataset which was prepared using WorkforceSim ver. 0.3.15. The dataset comprises more than 400000 records of 18 months' worth daily performance and attrition data for a factory of 508 workers with 35 columns. Due to employee turnover, there are data of a total of 687 workers and the observations cover regular and special daily events related to employees.
Data Source - https://www.kaggle.com/datasets/gladdenme/factory-workers-daily-performance-attrition-s

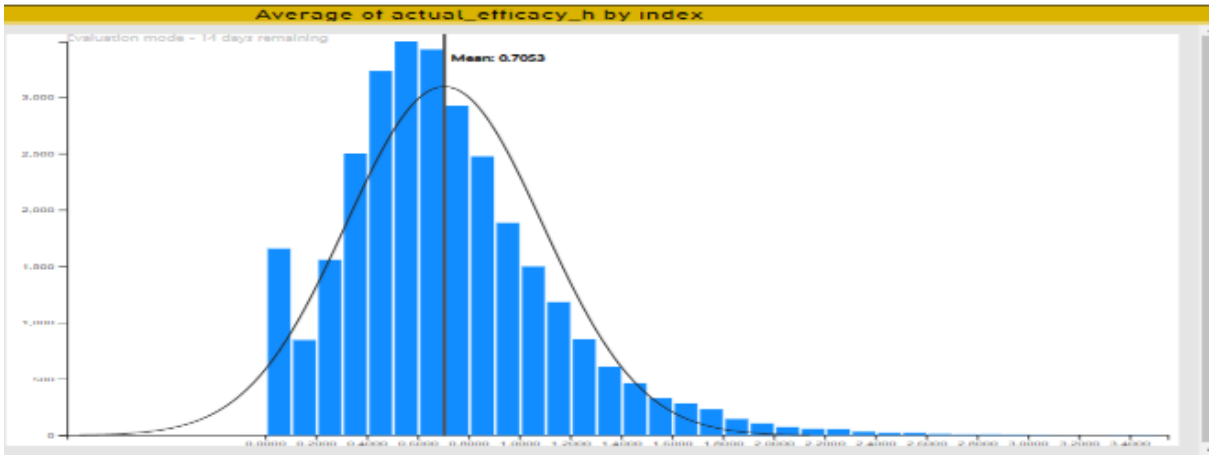
| | |
|---|---|
| sub ID | Employee ID number for the worker who is the subject of the event reflected in a row. |
| sub fname | The first name of the event's subject. |
| sub lname | The last name of the event's subject. |
| sub age | The age of the event's subject. |
| sub sex | The sex ("M" or "F") of the employee who is the event's subject. |
| sub shift | The name of the shift on which the event's subject works. |
| sub team | The name of the team on which the event's subject works. |
| sub role | The organizational role of the employee who is the event's subject. |
| sub coll IDs | A list containing the ID numbers of the other employees who are immediate co-worker-peers of the events. |
| sub colls same sex prtn | Portion of the subject's immediate co-worker-peers who are of the same sex. |
| sub health h | Health stat of the event's subject. |
| sub commitment h | Commitment stat of the event's subject. |
| sub perceptiveness h | Perceptiveness stat of the event's subject. |
| sub dexterity h | Dexterity stat of the event's subject. |
| sub sociality h | Sociality stat of the event's subject. |
| sub goodness h | Goodness stat of the event's subject. |
| sub strength h | Strength stat of the event's subject. |
| sub openmindedness h | Open-mindedness stat of the event's subject. |
| sub workstyle h | The subject's "Workstyle group" influences his or her Efficacy |
| Sup ID | The employee ID number of the person who was supervising the event's subject at the time when the event occurred. |
| sup fname | The first name of the subject's supervisor |
| sup lname | The last name of the subject's supervisor |
| sup age | The age of the subject's supervisor |
| sup sub age diff | An integer representing the age of the event's subject subtracted from the age of the person who was supervising that |
| sup sex | The sex ("M" or "F") of the subject's supervisor |
| sup role | The organizational role of the subject's supervisor. |
| sup commitment h | Commitment stat of the subject's supervisor. |

| | |
|---|---|
| sup perceptiveness h | Perceptiveness stat of the subject's supervisor |
| sup goodness h | Goodness stat of the subject's supervisor. |
| event date | The date of the event |
| event week in series | An integer representing the one-indexed week of the simulated time period |
| event day in series | An integer representing the one-indexed day within the simulated time period |
| event weekday num | A zero-indexed integer corresponding to the day of the week |
| event weekday name | The English-language name of the day of the week |
| Efficacy | Level of Efficacy that was displayed by the subject on the given day |

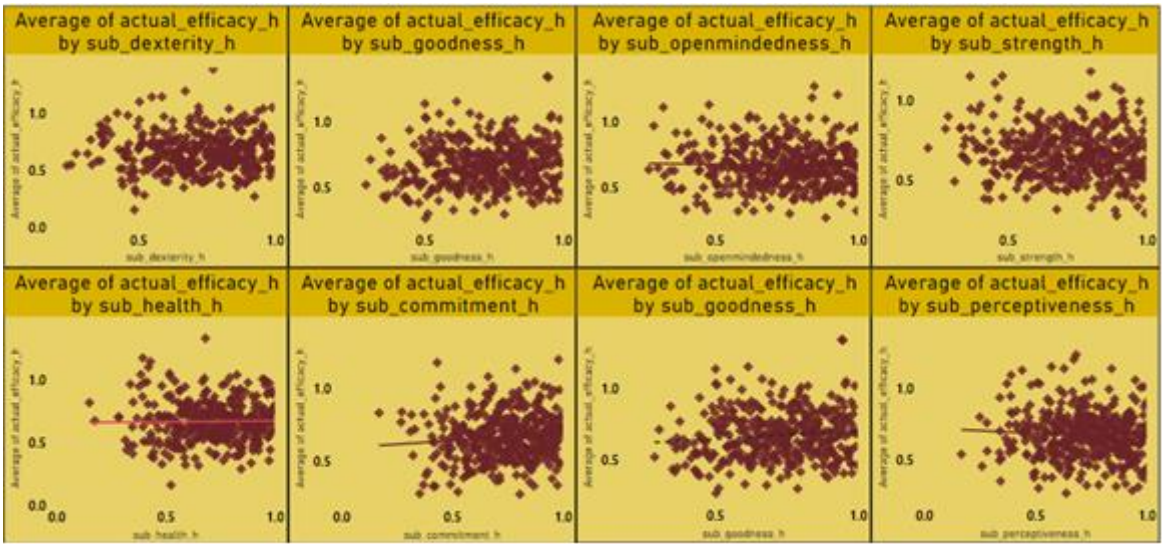
## Methodology

- Initially the dataset consisted of 411948 records with 42 columns, since our main objective was to find the factors affecting a worker's efficiency, using excel filtering option 'Efficacy' was filtered from 'record_comptype' column.
- Only 191657 records were of worker's efficacy. Among the workers there was a production director 'Robert Grigoriyan' who was not assigned to any team, group or supervisor.
- So, another 385 records of him were removed from the dataset and the rest 191272 records were uploaded to python environment for further cleaning and pre-processing.
- There were no missing values or duplicates and unnecessary columns were removed in order to obtain a clear dataset and the final dataset consisted of 191272 observations and 24 columns.
- Pre-processed datafile was exported power bi in order to do exploratory data analysis and necessary histograms, pie charts, bar graphs and scatter plots were drawn using power bi.
- Even some boxplots were drawn using seaborn library in python for our exploratory data analysis.
- A multiple linear regression model was fitted in Rstudio.
- Catogorical variables was converted to factor and data were split into train set and test set and a model was fitted using step AIC (direction = both) function.
- Since we are supposed to fit machine learning models to obtain most accurate model, following supervised machine learning models were fitted respectively.
- Ridge and Lasso models were fitted to the total dataset and root mean square error was obtained separately.
- A regression tree was fitted to the all variables and the obtained tree was pruned to obtain more accurate results.
- Further a Random Forest model was fitted using python with the aim of obtaining a precise model.
- After that Neural Network algorithm was fitted to the dataset.

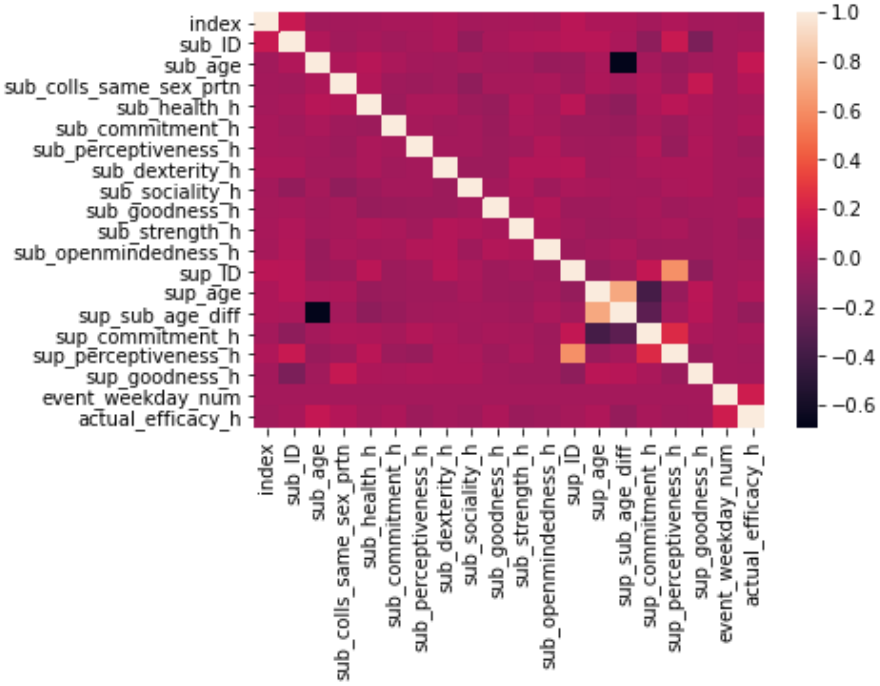
## Exploratory Data Analysis



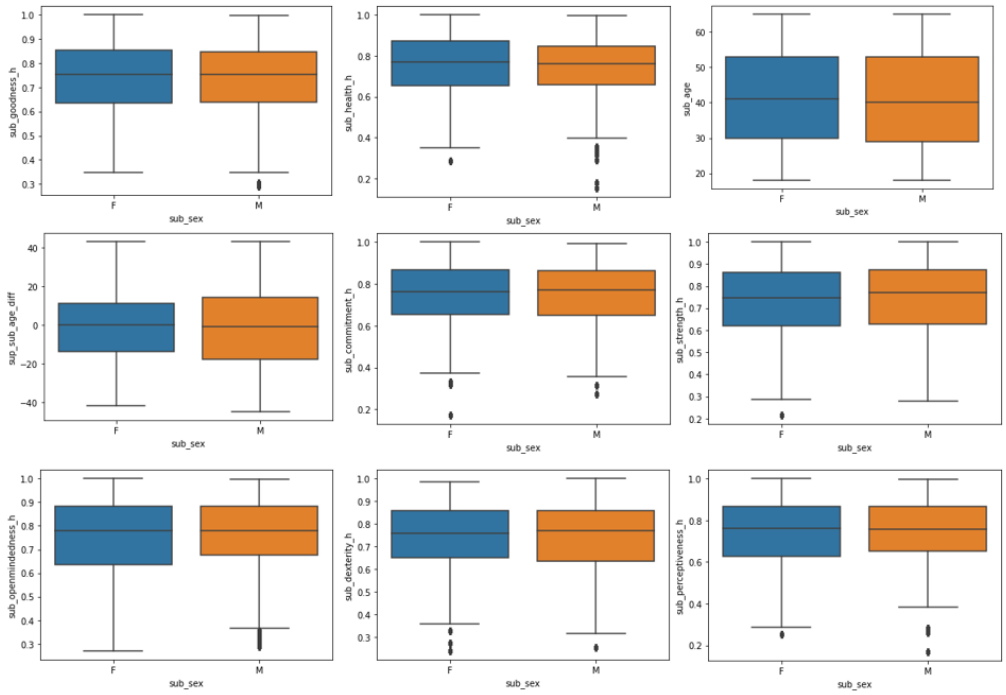

The above histogram shows the nature of the distribution of response variable, actual efficacy. It seems to be slightly positively skewed. But it can be ignored since its tail was not so long and shape had not a big difference with the symmetrical bell shape.




The above figure shows the correlation between the response variable (actual efficacy) and another 08 continuous variables as worker's dexterity, worker's goodness, worker's openmindness, worker's strength, worker's health, worker's commitment, worker's sociality and worker's perceptiveness.
All plots didn't show any considerable correlation with actual efficacy (response variable).




Above heat map shows the correlation of all variables in the data set and we can see that only sup_age and sup_sub_age_diff have a correlation of about 0.6 while very strong negative correlations are observed in sup_age vs sup_commitment_h and sub_age vs sup_sub_age_diff.




Here are the boxplots drawn for worker's goodness, health, age, commitment, strength, open mindedness, dexterity, perceptiveness and sup_sub_age_diff with gender of the worker. We can see that the scores are given from 0 to 1 for workers qualities and even though the boxplots show outliers, they can be neglected because those are still between the scoring system. In both male and female distributions, we can see similar variations in goodness, age, sup_sub_age_diff, commitment and strength and slight variations in health, open mindedness, dexterity and perceptiveness.

# Model Fitting

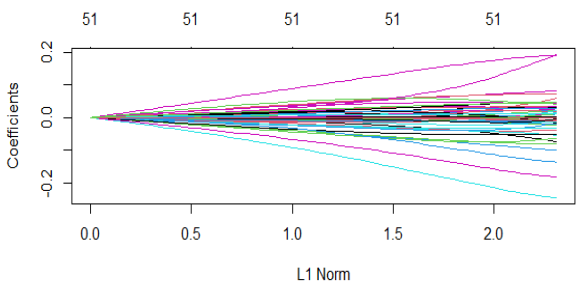
The final model taken from Multiple Linear Regression

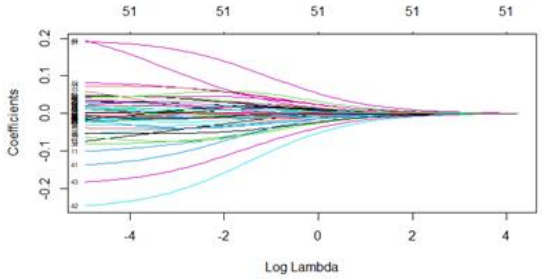
lm(formula = actual_efficacy_h ~ sub_age + sub_sex + sub_shift + sub_team + sub_role + sub_colls_same_sex_prtn + sub_health_h + sub_commitment_h + sub_perceptiveness_h + sub_dexterity_h + sub_sociality_h + sub_goodness_h + sub_strength_h + sub_workstyle_h + sup_age + sup_sex + sup_commitment_h + sup_perceptiveness_h + sup_goodness_h + event_weekday_num, data = train)

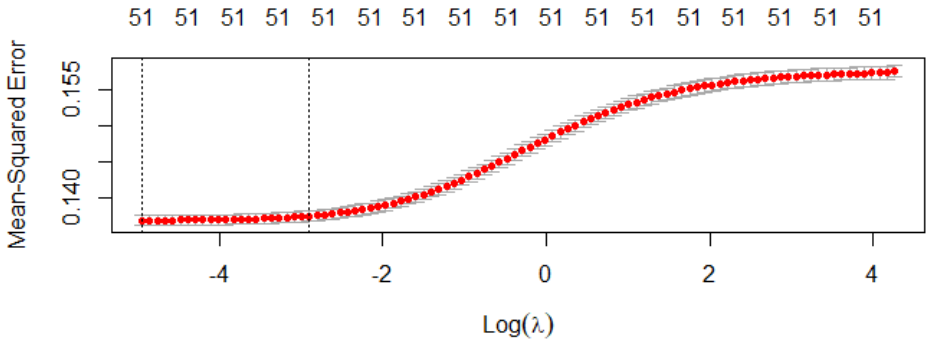
| | | |
|---|---|---|
| Residual standard error | : | 0.37 on 152969 degrees of freedom |
| Multiple R-squared | : | 0.1309 |
| Adjusted R-squared | : | 0.1307 |
| F-statistic | : | 490.4 on 47 and 152969 DF |
| p-value | : | < 2.2e-16 |

Since the p-value is less than 0.05, the model is significant, but the adjusted R-squared value is 0.1307. Adjusted R-squared value is used to compare the goodness-of-fit for regression models that contain different numbers of independent variables. Therefore, the predictor variables for the multiple linear regression model explained 13.07% of the variation in the response variable, the actual efficacy of the worker. This is not a good fit. Therefore, another model was tried out.

Ridge Regression

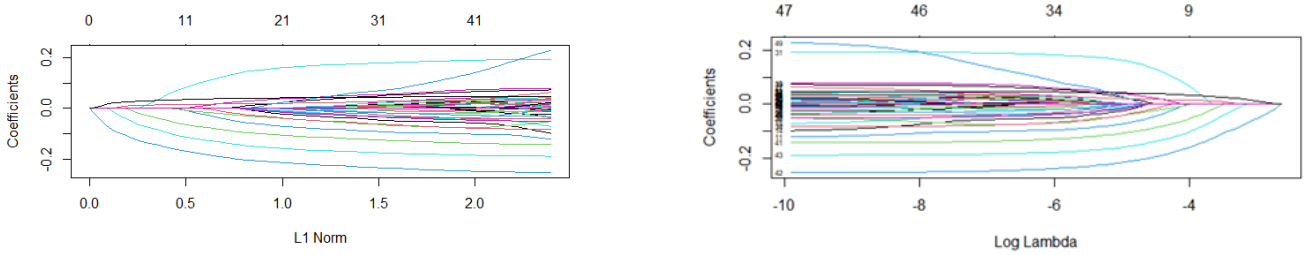


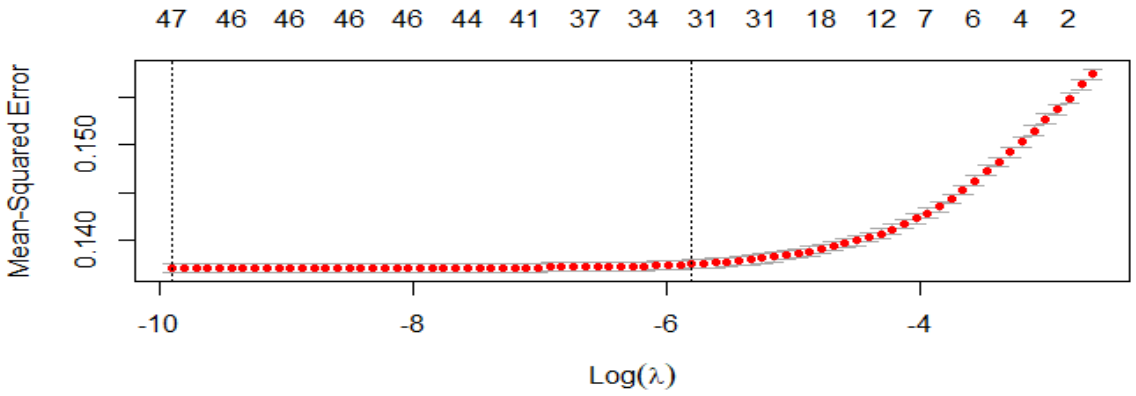



In order to get a more accurate model, it was used ridge and lasso regression techniques.
According to the coefficients plots obtained under the ridge model, it can be said that the expected outcome was gained. Because the coefficient lines seem to be converged to the zero line
The minimum lambda value was 0.00712 according to the plot of mean squared error vs log(lambda).
The final outcome, i.e. mean squared error was 0.13754
The standard error of MLR was 0.37. Therefore, ridge model was better than the MLR model.

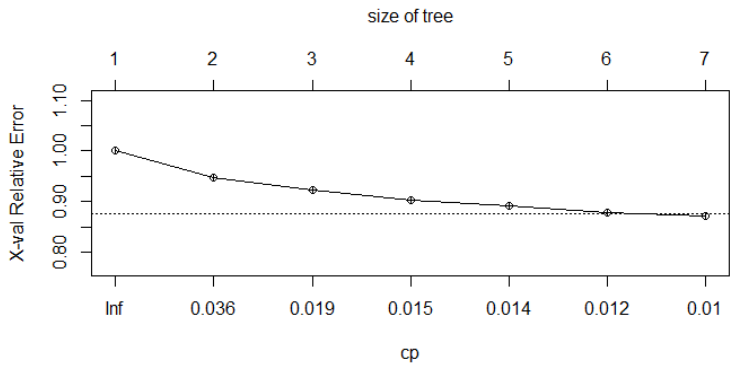
Lasso Regression



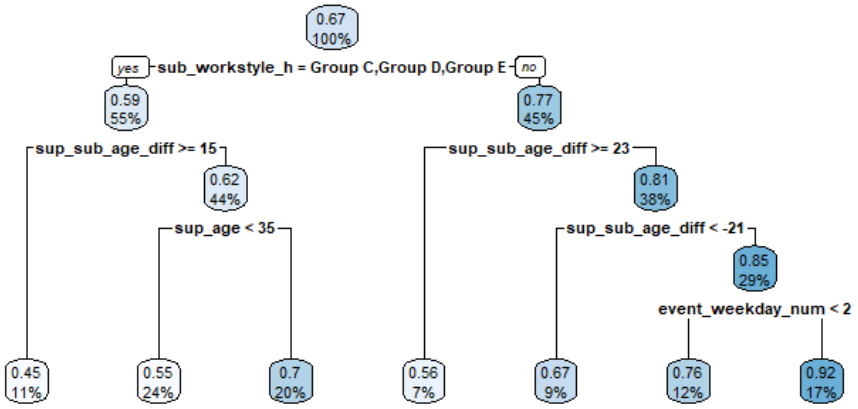

In the lasso model, the received mean squared error was 0.1373.
Coefficients plots show better convergence of coefficients to zero line. The minimum lambda value was 0.00005.
The mean squared error of lasso model was little bit less than mean squared error of ridge model.

Regression Tree



Complexity Parameter of the tree (cp) is 0.01. It provides the optimal pruning based on the cp value. We pruned the tree to avoid any overfitting of the data.

**Pruned Tree**




MSE value is 0.1350135.

Random Forest

When number of estimators are 20, following were the metrics obtained

Mean Absolute Error: 0.26474401904824674

Mean Squared Error: 0.13077862124041184

Root Mean Squared Error: 0.36163326899002507

When number of estimators are 200, following were the metrics obtained

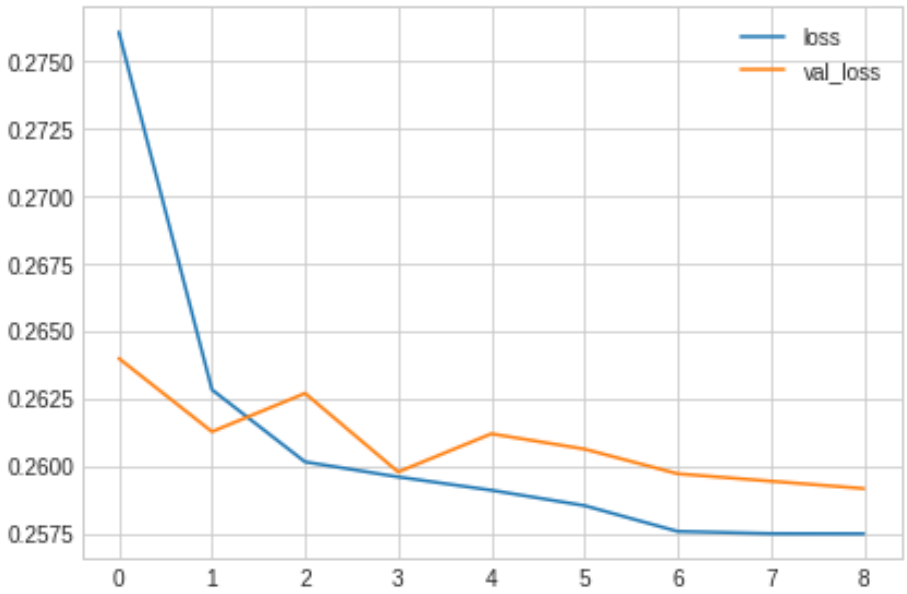
Mean Absolute Error: 0.26451310805620193

Mean Squared Error: 0.1305081129441407

Root Mean Squared Error: 0.36125906624490506

Even though the random forest model is fitted after increasing the number of estimators by 10 times, MSE and RMSE have only changed slightly but less than the previous values.

Neural Network



According to the above plot, it can be said that the neural network model was not overfitted. The final obtained mean squared error was 0.12821.
Out of all mean squared errors obtained, the best least value was the value that was gained by the neural network model.

## Conclusions

- Since R2 value of MLR is low and residual standard error is 0.37, it was necessary to fit another model to achieve high accuracy.
- Then Ridge and Lasso regression was performed and obtained MSE values were 0.1375 and 0.1373 respectively. So, models obtained through Ridge and Lasso regression were better fits than the MLR and Lasso model was the best out of the two.
- We then decided to apply decision trees. Since the dependent variable is quantitative, a regression tree was fitted and the obtained MSE value was 0.1350 which is lesser than earlier.
- Again, Random Forest was applied to the dataset with the motive of obtaining higher accuracy and the MSE obtained with 20 estimators was 0.13078 and MSE with 200 estimators was 0.1305. So, the model with highest number of estimators has the lowest MSE with more accuracy.
- Finally, Neural Network model was fitted and the MSE obtained was 0.1282 which gives the least MSE out of all models fitted and the most accurate model is the Neural Network model.