# **Documentation**

* There are 4 categorical features and 8 numeric features in the data.
* There are outliers & exponentially distributed data in output “area” feature. So applied logarithm transformation to make it normalized. Also, I experimented with outliers with the boundary cutoff (replacing all over the 80th percentile area to the 80th percentile). However, the upper bound cutoff transformation resulted in more error. Hence, I rejected the upper bound cutoff transformation approach.
* The outliers in the input feature were found using the Z-score, which is over 3 and replaced those values with the median values.
* Further, in pre-processing, we encoded the categorical features to one-hot encoded values.
* We created one more categorical feature to separate the fired or not. We built the SVM, Logistic Regression model to predict this feature. Finally, the results were produced with the help of logistic regression, as the accuracy was better.
* We experimented with Linear Regression (with Lasso, Ridge), SVM, Decision Tree, and Random Forest to predict the area feature.
* Each model has experimented with a nested validation method. Also for SVM, Decision Tree and Random Forest models, we visualized RMSE accuracy over the different parameters.
* 22 Out 42 Features coefficient is 0 in the Lasso model. So we applied Ridge Regularization to punish the coefficient instead of change to 0. However, the ridge regression increased the error than the lasso.
* SVM has a similar RMSE as the lasso model, but R2\_Score was better than all other models.
* Finally, the Decision Tree makes the best RMSE score. However, if we consider the R2\_Score metric SVM model is the optimal selection for this data.