# Documentation

* There are 4 categorical feature and 8 numeric feature.
* There are outliers & exponential in output "area" data. So applied logarithm and experimented outliers with upper bound cutoff. But the upper bound cutoff selection is not giving the optimal result. Hence I did not proceed with upper bound cutoff.
* Also, find the outliers in input feature using the Z-score which is having more than 3 and replaced those values with the median values.
* Further, in preprocessing, we encoded the categorical one hot encoded feature
* We created one more categorical feature to say whether it is burned or not based on the input variables. We built SVM, Logistic Regression to predict this metric. Finally, produced the result with Logistic regression as it is better accuracy.
* Then we proceeded to predict the area feature. We experimented with Linear Regression (with Lasso, Ridge), SVM, Decision Tree, and Random Forest.
* Each model experiments with nested validation method. Also for SVM, Decision Tree and Random Forest models, we visualized RMSE accuracy over the different parameter.
* Lasso is zero out the 22 features. So we applied Ridge Regularization. However, the ridge regression results in more error than lasso.
* SVM is having similar RMSE as lasso model but R2\_Score is better than all other models.
* Finally, the Decision Tree is giving the best RMSE score. But if we consider the R2\_Score metric Random Forest model is the optimal selection for this data.