# Predicting the efficiency of a mixer

By - Utkarsh Vardhan

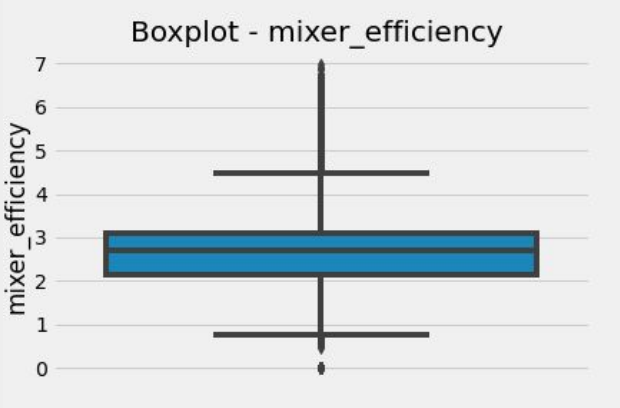
# Objective

1. Understand Rubber Mixer data and build a model capable of predicting the efficiency of a mixer.
2. Identify Features which will have maximum impact on the mixing cycle time.

# Target Data Statistic

1. 90% of the Mixing Efficiency is within the range 0 and 3.66.
2. Are higher values such as above 5 expected or can be treated as an outlier.
3. In case of such low distribution of higher values, we can look at the following options:
   a. Cap the Mixing Efficiency between a certain range.
   b. Choosing evaluation metric such as Root-Mean-Squared-Error (RMSE) between the logarithm of the predicted value and the logarithm of the observed value to ensure errors in predicting higher mixing efficiency and lower mixing efficiency will affect the result equally.

Mixing Efficiency




| Quantile | Value |
|---|---|
| 0.00 | 0.00 |
| 0.10 | 1.64 |
| 0.20 | 2.06 |
| 0.30 | 2.36 |
| 0.40 | 2.59 |
| 0.50 | 2.71 |
| 0.60 | 2.82 |
| 0.70 | 3.02 |
| 0.80 | 3.17 |
| 0.90 | 3.66 |
| 1.00 | 6.97 |

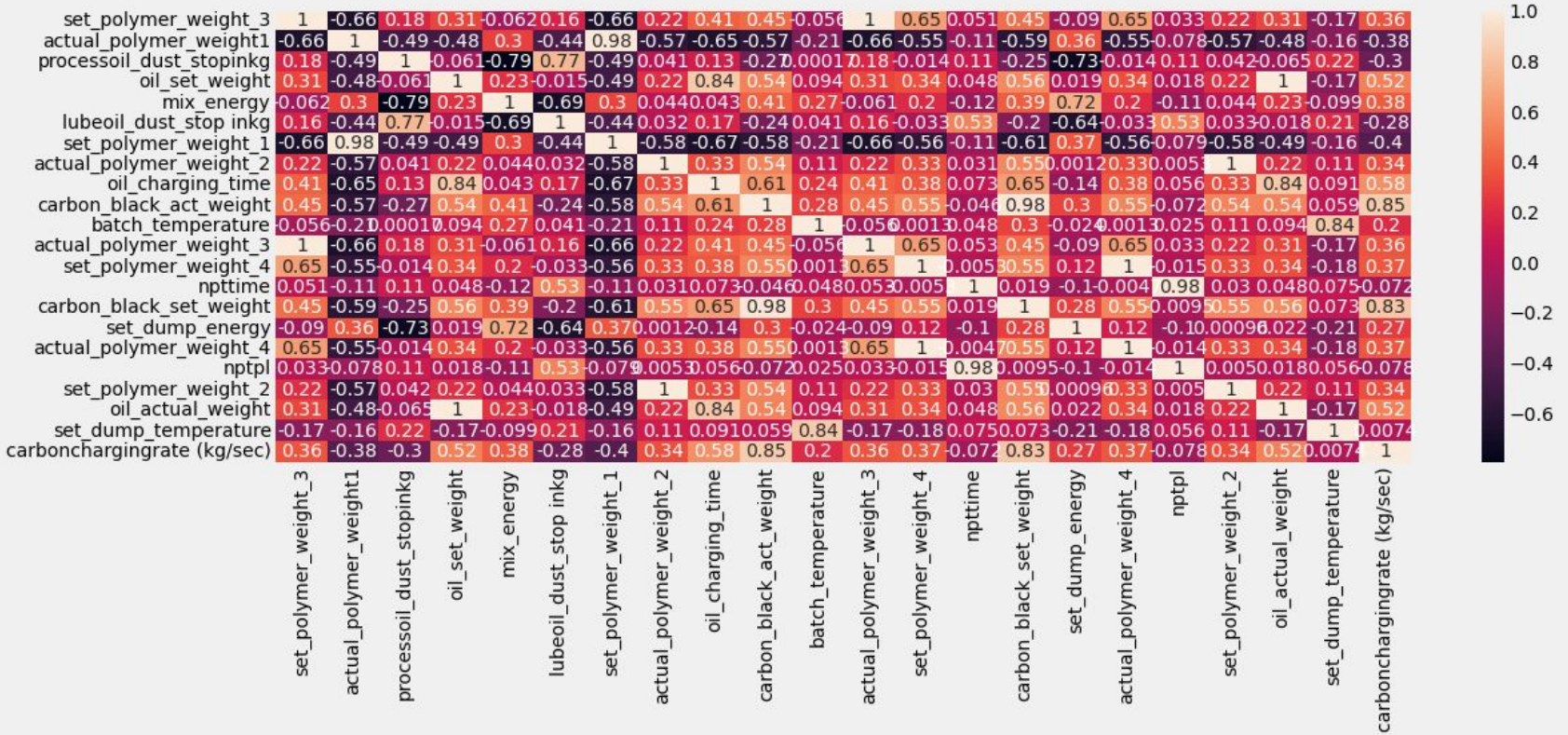
# EDA Insights

1. Features with no distinct values

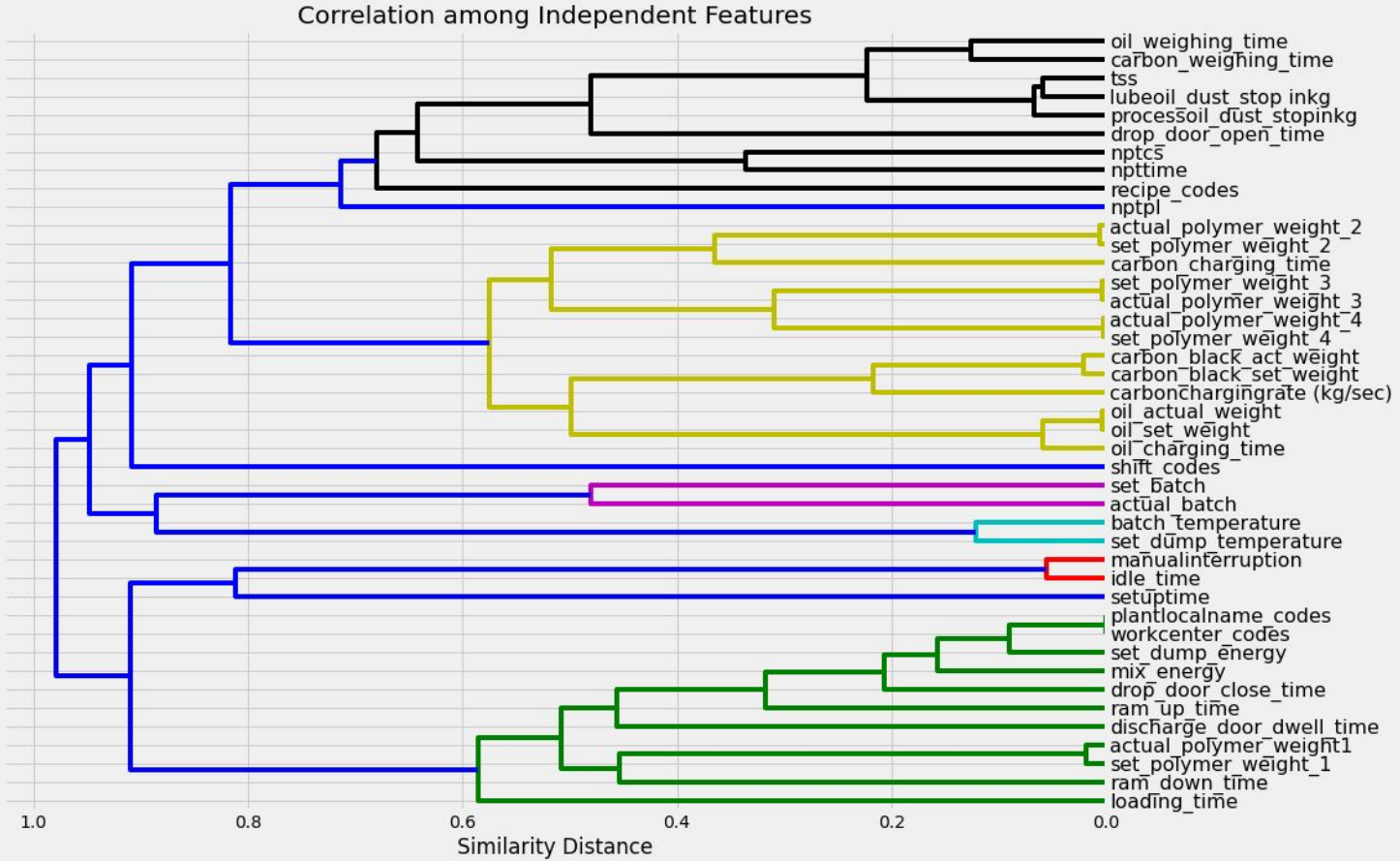
   'order_number', 'rotortype', 'remarks', 'batchtype', 'silica_set_weight', 'silanization', 'idle_energy', 'mixer_batch_completion', 'actual_rework_weight_4', 'set_rework_weight_4', 'zincoxide_actual_weight', 'zincoxide_set_weight', 'chemical_set_weight', 'chemical_actual_weight', 'silane_set_weight', 'silane_actual_weight', 'silica_charging_time', 'silica_weighingtime', 'silane_charging_time', 'mills', 'silica_actual_weight', 'plant', 'planttype', 'mixervfd'

2. For WorkCentre “P1002_W0107”,”lubeoil_dust_stop inkg”,”processoil_dust_stopinkg”,”oil_weighing_time”,”carbon_weighing_time” are all 0’s which is not the case with WorkCentre “P1002_W0105”.
3. For WorkCentre “P1002_W0105”,set_dump_energy is all 0’s which is not the case with WorkCentre “P1002_W0107”
4. For WorkCentre “P1002_W0107”,”tss” has value just 0 whereas “P1002_W0107” “tss” has values 1 and 2.
5. “Manualinterruption" and “idle time” are highly correlated ,as Manualinterruption =0 is always linked to idle time =0.
6. "Plantlocalname_codes" and “workcenter” are duplicate features.
7. Actual_polymer weight related features and Set_polymer weight related features has high correlations.
8. Npttime and nptpl has high correlation.
9. Mixer Efficiency differs for different recipes.
10. Avg Mixer Efficiency differs between different plant locations and could be because of different recipes used at each plant.
11. Mixing Cycle Time has higher correlation with the following numerical features.
    a. mix_energy
    b. Nptpl
    c. Npttime
12. Avg Mixing Cycle time differs for different recipe type.

# Correlation Plot



# Correlation Plot(contd)




*Lower the similarity distance higher is the correlation*

# Modeling Strategy:

1. Created Dependent Feature Mixing Efficiency - Total Weight/Mixing Time.
2. Sorted “localizedtimestamp” column in the dataset to create train,val and test set.
3. Dropped columns which has number of distinct values < 2 as they would not add any predictive power.
4. Encoded categorical features.
5. Ignored mixing_cycle_time as it has higher correlation with the target variable and could be a leaky feature.Can be discussed with the data team whether to consider it as an independent feature.
6. Dropping Duplicate Features like “manualinterruption”,”plantlocalname_codes”.
7. Choosing and dropping Correlated Features based on its impact on training data evaluation metric like "set_polymer_weight_1","set_polymer_weight_2","carbon_black_set_weight","oil_set_weight"
8. Used Permutation Importance method to identify important features on the validation set and cross referenced with Feature Importance on the train data to identify noisy features.
9. Calculated Correlations between Feature values between train and validation set.Few features like “carbon_black_act_weight”,”carbonchargingrate”,”carbon_charging_time” has very low correlations.Check with the data team to identify the nature of these fields and its dependency on other fields.
10. Nested Cross Validation to do Algorithm Selection and Hyperparameter Tuning of the algorithm to find the best algorithm and its hyperparameters.
11. Evaluate the model on the validation and test set.
12. Test model stability under different scenarios.

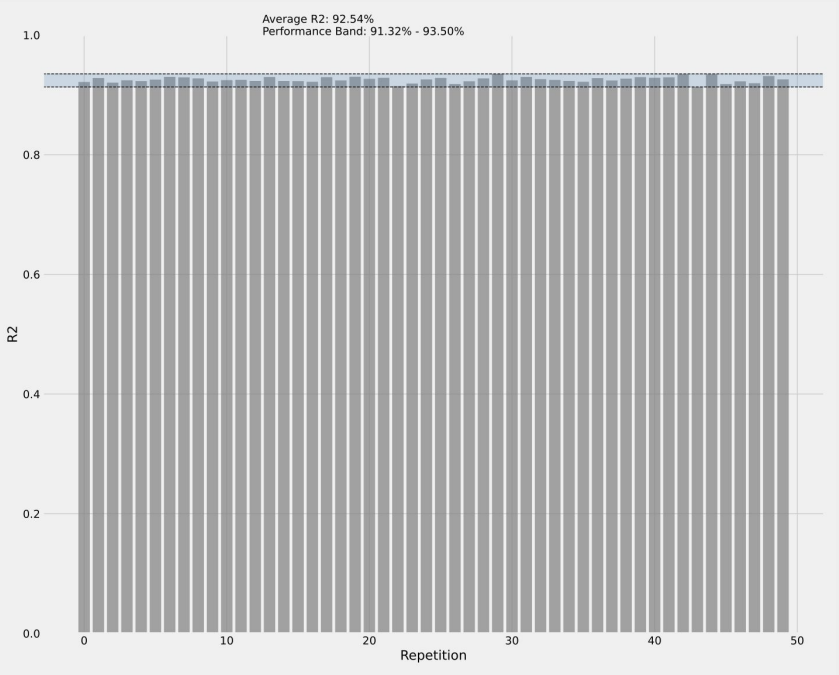
# Model Report Card

1. Dataset
   a. Train Set - 8998 samples between (Timestamp('2022-04-01 07:01:06.530000'), and Timestamp('2022-04-13 15:44:09.923000')
   b. Validation Set - 1000 samples between (Timestamp('2022-04-13 15:44:59.427000'), Timestamp('2022-04-14 22:10:24.810000')).
   c. Test Set - 1000 samples between (Timestamp('2022-04-14 22:12:47.850000'), Timestamp('2022-04-17 10:35:38.317000')).

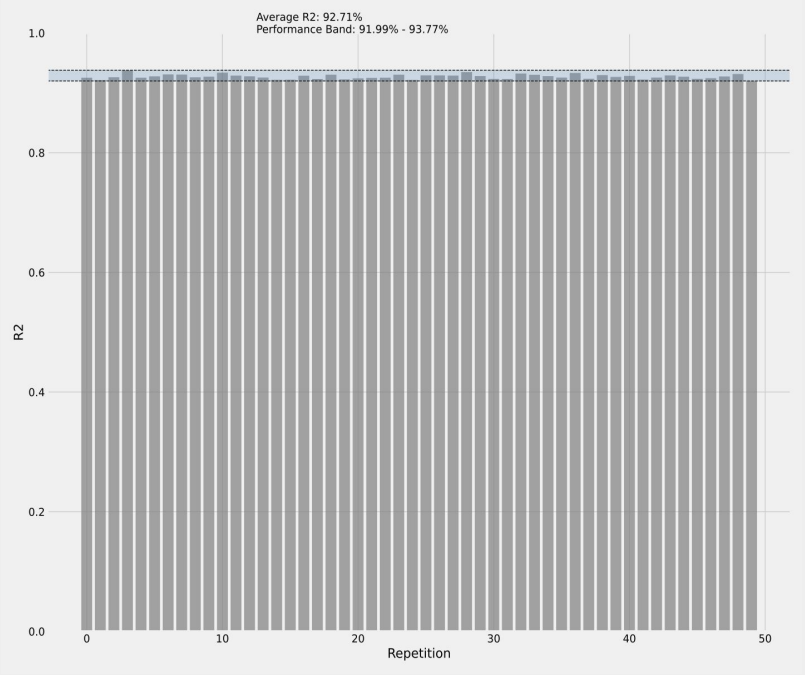
2. Features
   a. 13 Features used in the model:
   'mix_energy', 'ram_down_time', 'actual_polymer_weight1', 'recipe_codes', 'npttime', 'nptcs', 'set_dump_temperature', 'carbon_black_act_weight', 'batch_temperature', 'drop_door_open_time', 'oil_actual_weight', 'actual_polymer_weight_2', 'processoil_dust_stopinkg'.
   b. Feature Transformations - None
3. Evaluation
   a. $R^2$ on validation dataset - 0. 978
   b. $R^2$ on test dataset - 0.975
4. Performance Estimates:
   a. Varying the sample size
      i. 50:50 split
         1. Average $R^2$: 0.925
         2. Performance Band: 0.913 to 0.935
      ii. 70:30 split
         1. Average $R^2$: 0.927
         2. Performance Band: 0.919 to 0.937

# Model Stability Reports

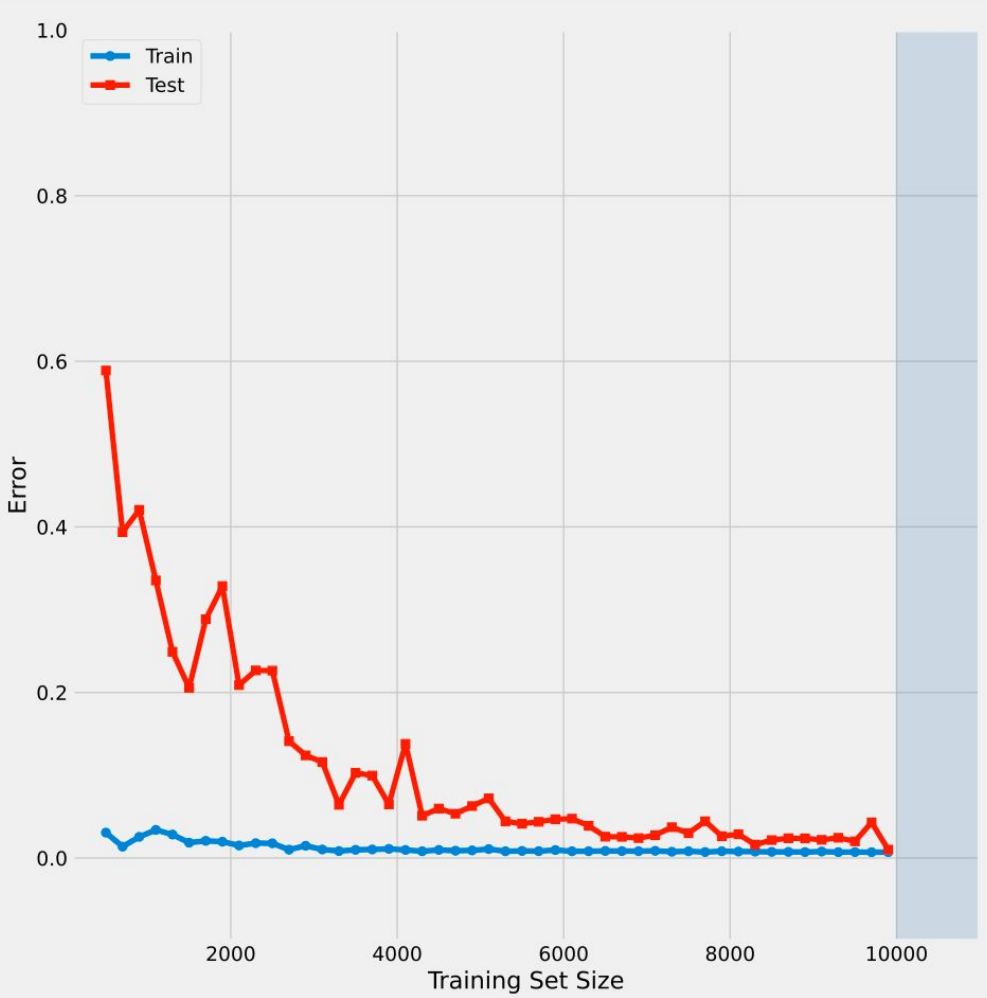
Train Test Split - 50:50




Train Test Split - 70:30



# Model Learning Curve



# Future Work

1. Divide the dataset into training,validation and test with no overlapping dates.
2. Identify interactions between the features and do feature engineering.
3. Identify the feature transformations of high cardinal features like “recipe” if encoded into the model.
4. Extensive Algorithm Selection -Exploration of both linear and non linear Black-Box Models.
5. Extensive Model Selection to identify the right set of features and hyperparameters.
6. Experiment on Building models catering to specific Plant Type or Plant Location/WorkCentre and check their performance.
7. Adding following features related to Date:
    a. Year
    b. Month of the year
    c. Week of the year
    d. Day of the week
    e. Month End ,Month Start etc
8. Transform the regression problem into a classification problem by binning the Mixing Efficiency values into makeshift classes.

# Evaluation Metric

Currently Used:

1. RMSE between the predicted value and the observed value.
2. R2

Possible Alternatives:

1. Root-Mean-Squared-Error (RMSE) between the logarithm of the predicted value and the logarithm of the observed value to ensure errors in predicting expensive houses and cheap houses will affect the result equally.
2. Clipping of the target values and using RMSE between the predicted value and the observed value.