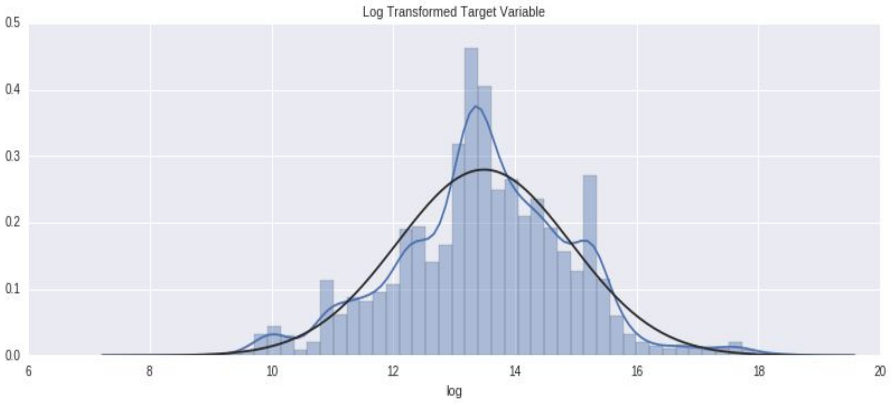
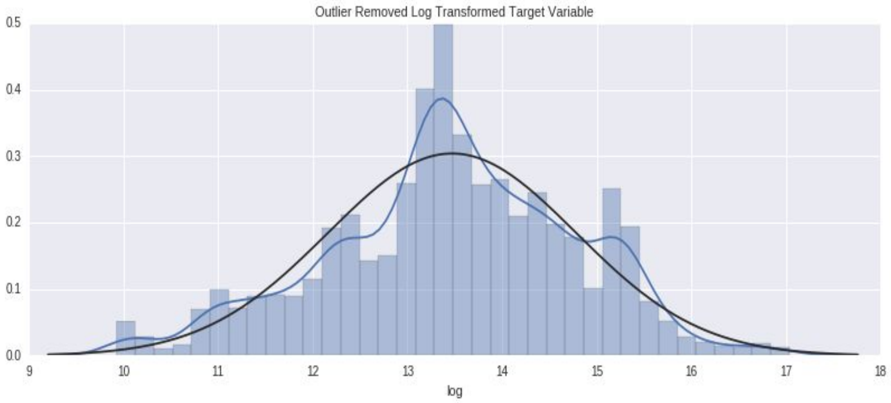
Rinse Over Run Model Analysis

prabod_cse

Preprocessing

- Replace negative sensor readings by 0 (supply_flow, supply_pressure, return_turbidity, return_flow and return_conductivity)
- Outlier Removal based on log transformed target variable (~200 outliers)
  - mean - 2.5 * std < log(target) < mean + 2.5 * std
- Training Dataset created to match the Testing Dataset using cleaning recipe to find missing value percentages.





Feature Engineering

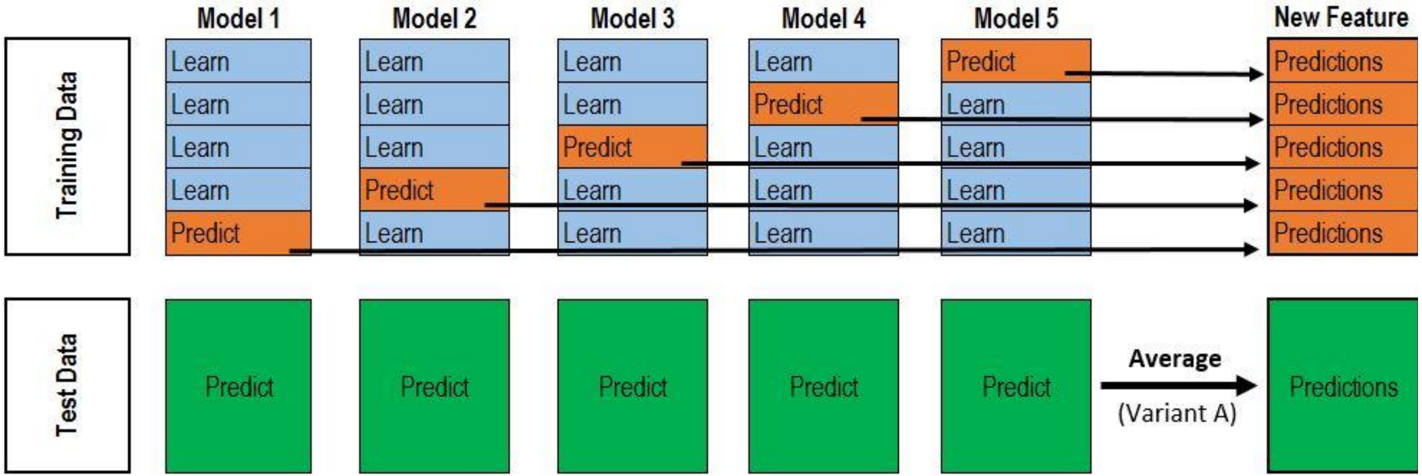
- **Summarized Time Series columns for all available data (mean, std, min, max, median, sum)**
- Summarized Time Series columns for each available phase in a process (mean, std, min, max, median, sum, skew, kurt)
- Summarized Time Series columns for last n records (n = **5**, 10, 50, **100**, 200, 500)
- Object Id and and Pipeline as categorical variables
- Time spent on each phase
- **Count of boolean true values for boolean columns**
- Time of day and day of week of the process as cyclic features
- Recalculated conductivity values at 25°C
- Derived Features
    - Interaction features of return and supply numerical columns
        - Supply -> flow, pressure,
        - Return -> flow, conductivity, turbidity, temperature,
    - Object Residue => supply_flow - return_flow

****Bold features are not used in the final model**

Model

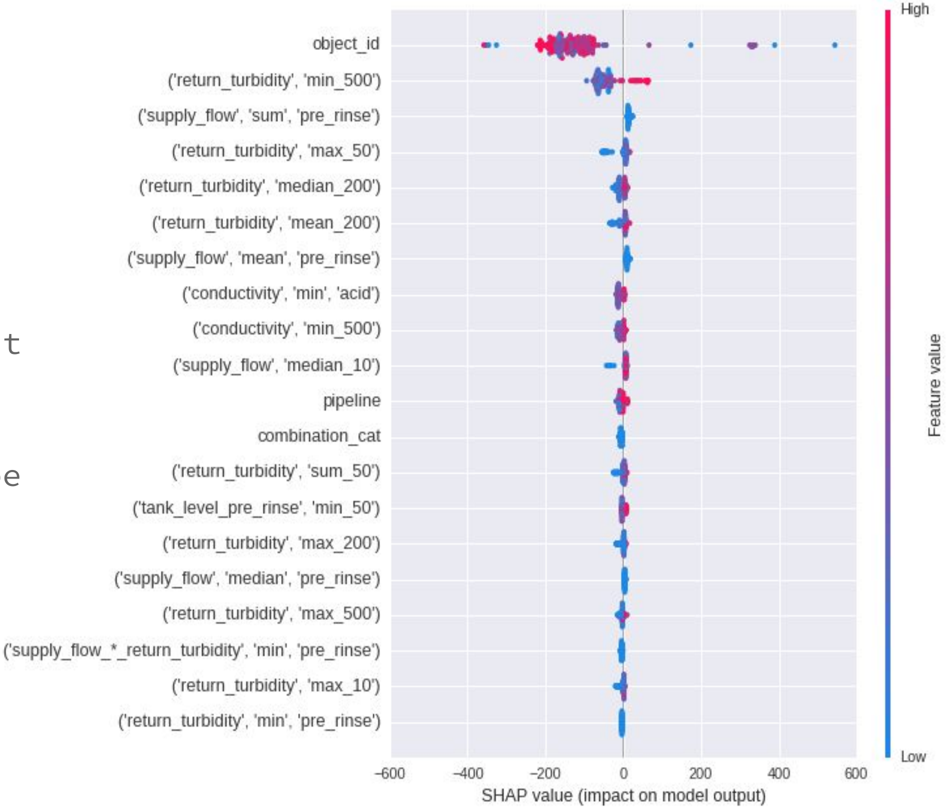
- Used stacked LightGBM model with 550 features (selected from 1100)
- Used a 10 fold stack
- Predicts the square root of the target variable

Local CV	Public LB Score	Private LB Score
0.2732	0.2826	0.2880



Sensitivity Analysis for Recipe (Pre -> acid -> Final)

- Model is highly dependent on object_id
- Higher min value for return_turbidity, conductivity, and pre_rinse_tank_level for last 500 records drives the model output higher and vice versa
- Lower supply_flow during pre_rinse phase drives the model output higher
- Higher max return_turbidity value during last 500,200 and 10 records drives the model output higher
- Model depends on the pipeline and recipe type
- Min conductivity value during acid phase controls the decrease of the model output
- Each of the 550 features contribute to the model at various instances of training data.
- For this recipe test data doesn’t have pre_rinse data. Therefore every prediction done with removing pre_rinse phase.



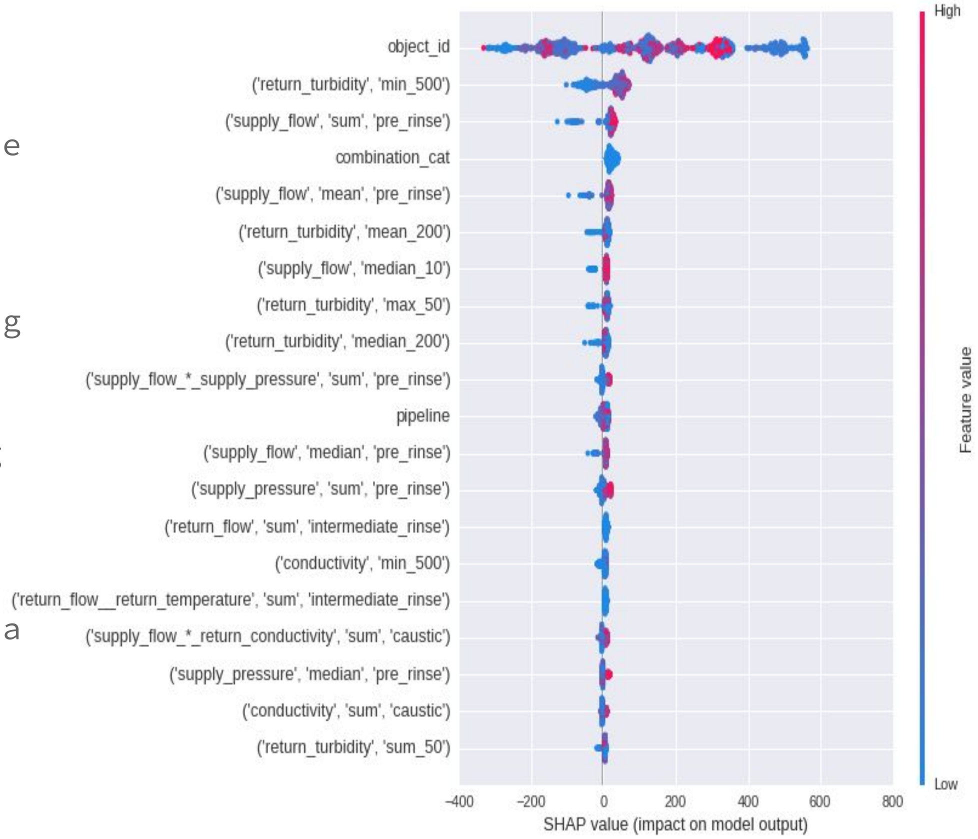
Sensitivity Analysis for Recipe (Pre -> acid -> Final)

Model predicts higher value	Model predicts lower value
	

- Model is responsive to last n records and able to adjust the output value accordingly. (n = 10, 50, 200, 500)
- By monitoring min values for last 500 records, we can determine the turbidity present in the final rinse cycle.
- About 95% of the instances of this recipe corresponds to a output value which is lower than the base value.
- Rest 5% is corresponds to the deviations from the cleaning recipe.

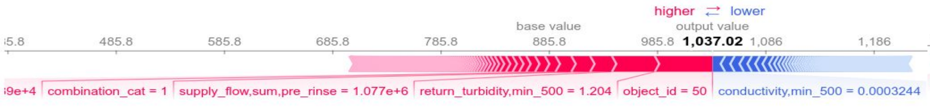
Sensitivity Analysis for Recipe (Pre -> caustic -> Final)

- Model is highly dependent on object_id
- Higher min value for return_turbidity, conductivity for last 500 records drives the model output higher and vice versa
- Higher supply_flow sum during pre_rinse phase drives the model output higher
- Higher supply pressure sum and median during pre_rinse drives the output higher
- Model depends on the pipeline
- Low return_flow * return_temperature during intermediate rinse increase the model output
- Higher Supply_flow * return_conductivity is resulting higher model output and vice versa



Sensitivity Analysis for Recipe (Pre -> caustic -> Final)

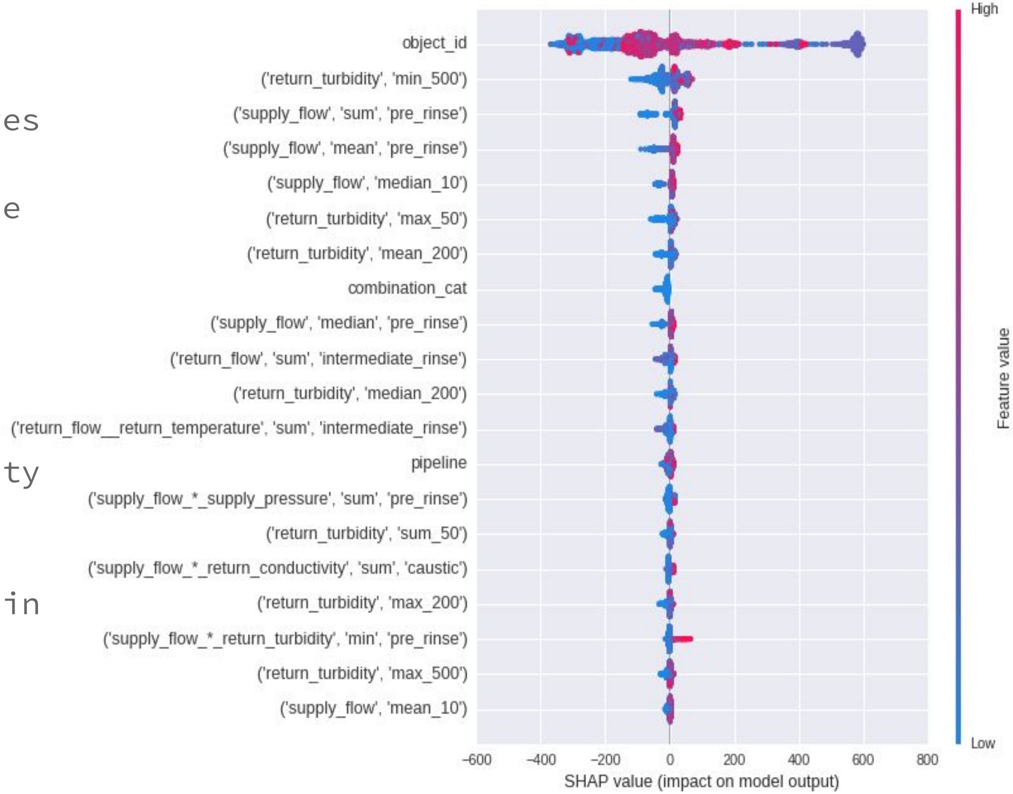
Model predicts higher value	Model predicts lower value



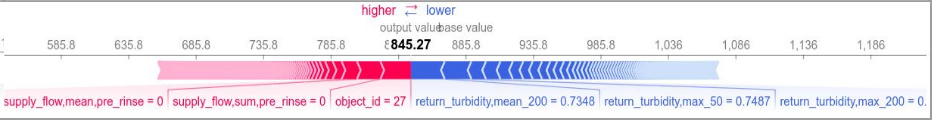
- Model is responsive to last n records and able to adjust the output value accordingly. (n = 10, 50, 200, 500)
- By monitoring interaction features and last n records we can predict the turbidity
- This recipe is mostly used in wineries and breweries
- About 85% of the instances of this recipe corresponds to a output value which is lower than the base value.
- Rest 15% is corresponds to the deviations from the cleaning recipe.
- It is given that occasionally these objects are sent through an acid phase which accounts for the deviations

Sensitivity Analysis for Recipe (Pre -> caustic -> inter-> acid -> Final)

- Model is highly dependent on object_id
- Higher min value for return_turbidity, conductivity for last 500 records drives the model output higher and vice versa
- Higher supply_flow sum during pre_rinse phase drives the model output higher
- Model depends on the pipeline
- Low return_flow * return_temperature during intermediate rinse increase the model output
- Higher Supply_flow * return_conductivity in caustic phase is resulting higher model output and vice versa
- Higher Supply_flow * return_turbidity in pre_rinse phase is resulting higher model output and vice versa



Sensitivity Analysis for Recipe (Pre -> caustic -> inter-> acid -> Final)

Model predicts higher value	Model predicts lower value
	

- Model is responsive to last n records and able to adjust the output value accordingly. (n = 10, 50, 200, 500)
- By monitoring interaction features and last n records we can predict the turbidity
- This recipe is mostly used in breweries and cleaning in dairy sector.
- About 90% of the instances of this recipe corresponds to a output value which is closer to the base value.
- Rest 10% is corresponds to the deviations from the cleaning recipe.

Conclusion & Future Works

- Model is highly dependent on Object Id, pipeline and recipe.
- Pre rinse phase is the most important phase to determine the turbidity in the final rinse.
- Supply_flow, supply_pressure, return_flow, return_conductivity, return_turbidity are the most important features.
- Interaction features are also plays a major role.
- Selection of performing features is important.
- Preprocessing of data led to a better score.
- More feature engineering would lead to a better solution.