Flipkart

GRiD 5.0

Problem statement title: Problem 2 - Drug Quality Test using AI Simulation

Team Name: 686250-UVH319L7

Team Details

Team Name	686250-UVH319L7		
Institute Name	International Institute of Information Technology Bangalore		
Team Members	1 (Leader)	2	3
Name	Brij Desai	Nitheezkant R	Yukta Rajapur
Batch	2026	2026	2026

Glossary

- **API (Active Pharmaceutical Ingredient):** An Active Ingredient is any ingredient that provides a biological effect in the diagnosis, cure, mitigation, treatment, or prevention of disease.

- **Excipient**: An inactive substance that serves as the vehicle or medium for a drug or other active substance.

- **Pre-production quality control:** This is a set of tests done by pharmaceutical manufacturers ***before starting the production of a batch****.* This included various tests on the raw API and Excipients.

- **Features:** In machine learning, features are individual independent variables that act like input into the system.

- **Labels**: In machine learning, labels are individual independent variables that act like the output of the system.

- **R-squared Score**: R-squared (R2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable in a regression model. An R2 of 1 indicates that the regression predictions perfectly fit the data.

Use-cases

1 Prediction of drug quality of the batch before starting production.

2 Lower wastage rate for pharmaceutical companies.

3 Higher quality and cheaper drugs for the public because of low wastage.

Problems to address

- Pharmaceutical manufacturers are required to conduct intensive quality control in each stage of the production process for each batch of drugs produced.
- When a particular batch doesn't meet the quality standards, it is required not to be sold in the market.
- Discarding a substandard batch is a difficult decision that manufacturers have to take because of the heavy losses that they will have to face.
- This sometimes leads the manufacturers to release drug batches of substandard quality into the market, especially in emerging economies and developing countries.
- It is difficult for regulatory authorities to set up labs to test the quality of drugs across the market.
- Hence addressing this problem at the root level (manufacturing) is more effective than the node level (after release into markets) solutions.

Efficacy and Efficiency of drugs

The efficacy and efficiency of a particular drug can be defined by the 1) amount of impurities present, 2) the dissolution of the drug in bodily fluids, and 3) the amount of residual solvents present.

Impurities

For a drug product, any component that is not a formulation ingredient is considered an impurity. The various sources of impurity in pharmaceutical products are – reagents, heavy metals, ligands, catalysts, other materials like filter aids, charcoal, and the like, degraded end products obtained during /after manufacturing of bulk drugs from hydrolysis, photolytic cleavage, oxidative degradation, decarboxylation, enantiomeric impurity, and so on.

Dissolution

The rate-limiting step in drug absorption from the gastrointestinal tract is drug dissolution from the dosage form. Drug API is intended to dissolve by a certain amount at a given time after intake. It is crucial for the sample to match this benchmark.

Residual solvents

Residual solvents in pharmaceuticals are defined as organic volatile chemicals that are used or produced in the manufacture of drug substances or excipients, or in the preparation of drug products. The solvents are not completely removed by practical manufacturing techniques.

Solution Statement

AI-based solution for quality control

Potential of AI

AI/ML models have the potential to process and learn from large amounts of quality control data. The models capture the relations between various features of the different quality-control tests and potentially predict the efficacy and efficiency of a given drug.

Should AI replace lab testing ?

No. Even the best AI/ML models are prone to errors, and there is very less room for errors in pharmaceutical drug testing. Importantly, AI/ML models learn only from data that it has been trained on. This means that there is no way for the model to predict the lapse in quality caused due to new factors. In such cases, a quality control protocol that relies on AI/ML models can prove dangerous by potentially falsely validating a substandard drug batch.
Hence all batches should definitely undergo thorough lab testing before release into markets.

Solution Statement

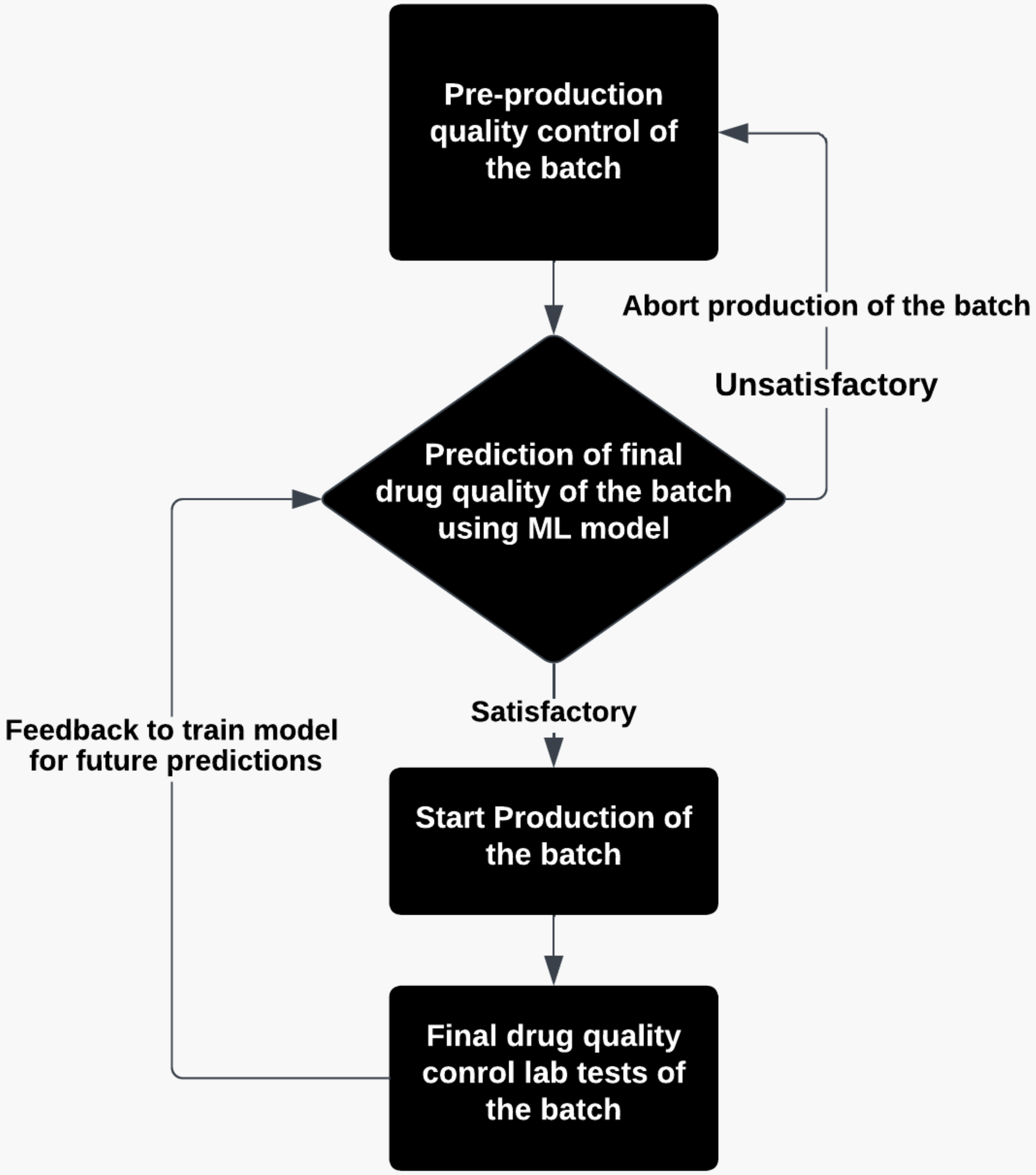
Leveraging AI Safely - Proposed Solution

The proposed solution enables manufacturers to predict the efficacy of the final drug before starting production of a batch using an AI/ML model that takes in the pre-production quality control data as input.

If the efficacy is not satisfactory, the production of the batch is not started, saving huge sums of money.

If the efficacy predicted is satisfactory, the production of the batch is started.

After production is complete, the lab tests are done to test quality. The result of the test is sent back to the model so that it trains itself. This establishes a continuous feedback mechanism for the model.


Pre-production quality control of the batch
Abort production of the batch
Unsatisfactory
Prediction of final drug quality of the batch using ML model
Satisfactory
Feedback to train model for future predictions
Start Production of the batch
Final drug quality conrol lab tests of the batch

Solution Statement

Dataset, Test Setup, Features and Models

Dataset and Test Setup

We are using one of the datasets taken from https://www.nature.com/articles/s41597-022-01203-x for the demonstration of the proposed idea. The dataset contains quality control data of 1005 batches of immediate-release, film-coated tablets. It contains information about the API and excipients (lactose, SMCC and starch). All data considered from the dataset is given in the next section.

In practice, there is **no additional test setup required, hence no additional .** Presently, almost all features used by the model are tested by the manufacturers. Manufacturers can even add features that are specific to them.

Features

Categorical features

- Family of the drug
- Strength (Dosage of API in mg)
- Batch size
- API type
- Batch number of API and excipients

Pre-Production lab test features

- Water content in API & excipients
- Particle size of API & excipients
- Impurities in API
- Percentage API content in raw material
- pH value, tap and bulk density of SMCC

Labels

1) Total impurities content in the final product (High-performance liquid chromatography Method)
2) Residual solvent content in the final product (gas chromatography method)
3) Average dissolution - the percentage of API released in 30 minutes.

Models

Instead of using a single model to predict efficacy, multiple models can be used. This gives more reliable information to the manufacturer.

Around 12 ML models with different hyperparameters were tested for predicting each of the 3 labels. The best ones for each of the labels were chosen.

The models chosen include:

- Gradient-boosted decision trees
- Bagged decision trees
- Gaussian process regression
- Support Vector machine

Results

- To demonstrate the accuracy, the best models suiting each of the labels have been simulated in MATLAB. These models have been trained with 70% of the data and tested on the rest 30%, with a 5-fold cross-validation.
- The plots for actual vs predicted labels have been shown in the following slides along with the R Squared Score.
- Given the less amount of data per kind of drug, the accuracy was found to be satisfactory to a great extent. Most of the points are seen to lie along the ideal prediction line.
- To demonstrate the deployment, a Python application has been built. A demo of this application is shown in the demo video attached to the submission.
- This application has models that have been trained with almost the whole dataset. A few data points have been set aside for the demo.
- This application shows exactly how manufacturers can use the proposed solution.

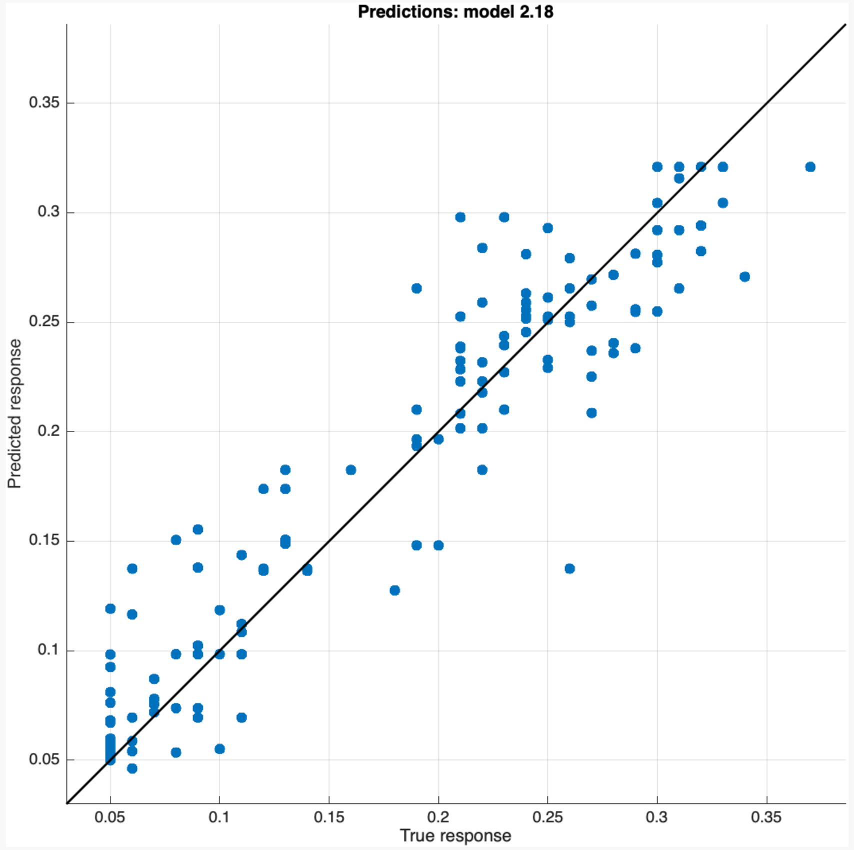
Total Percentage Impurity

The model name and the R-Squared Score of the test data are indicated below each figure.
The Black line represents the output of a perfect model.

Legend

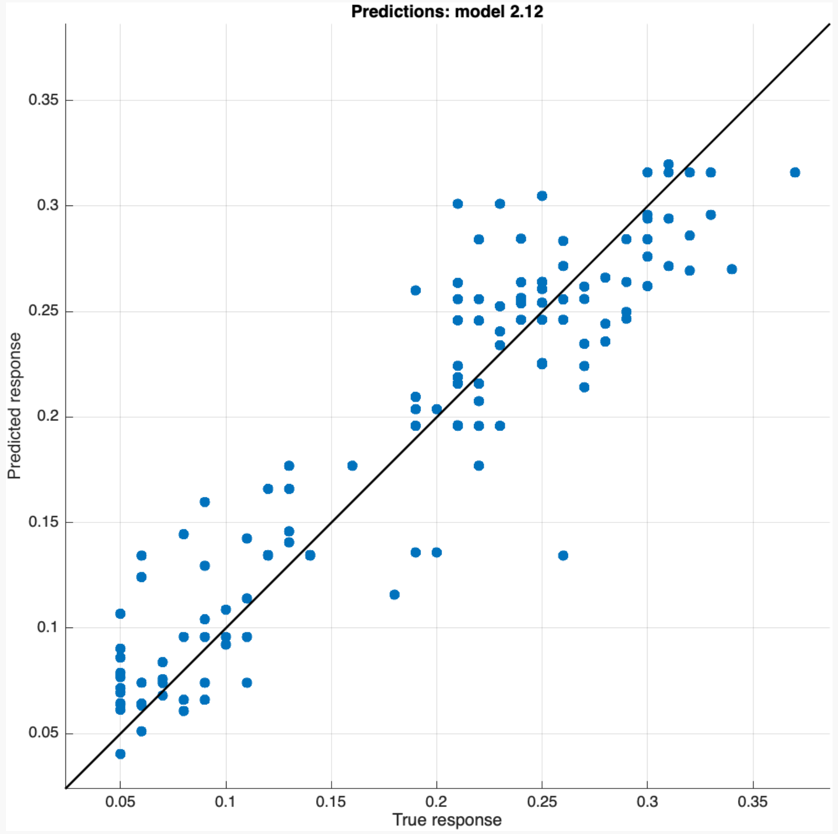
● Observations

— Perfect prediction



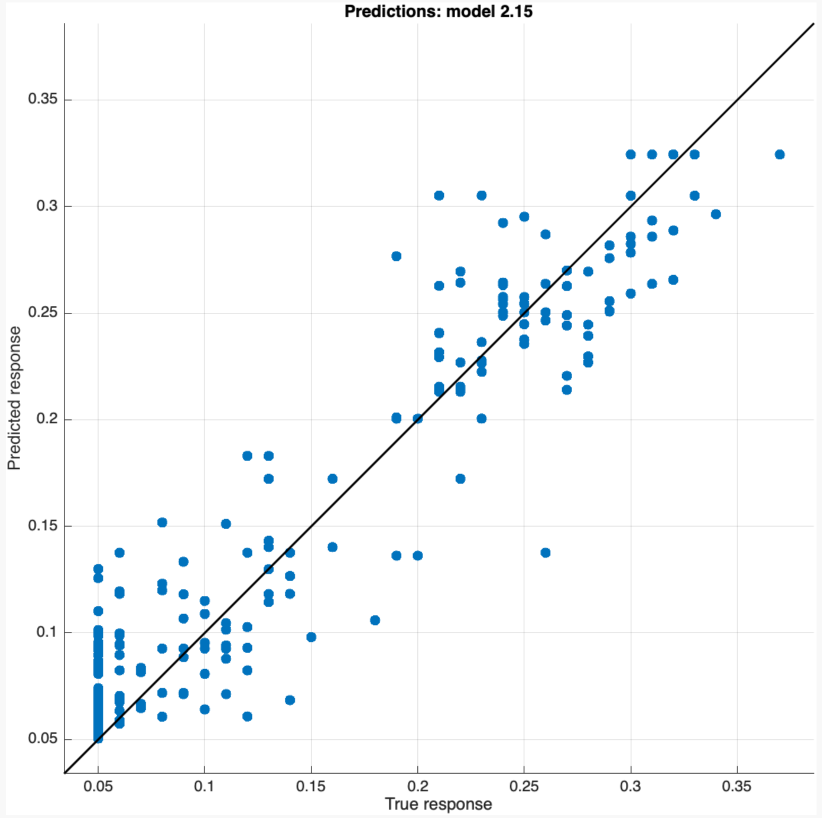
Gaussian Process Regression
(Exponential kernel)

0.91



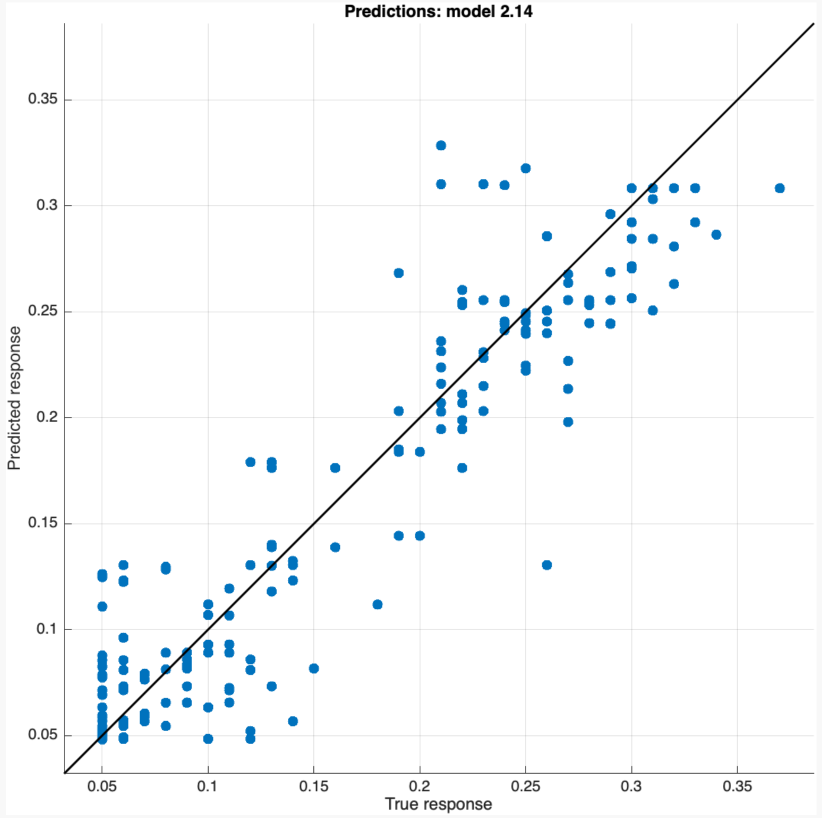
Support Vector Machine
(Medium Gaussian kernel)

0.90



Ensemble of Trees
(Bagged Trees)

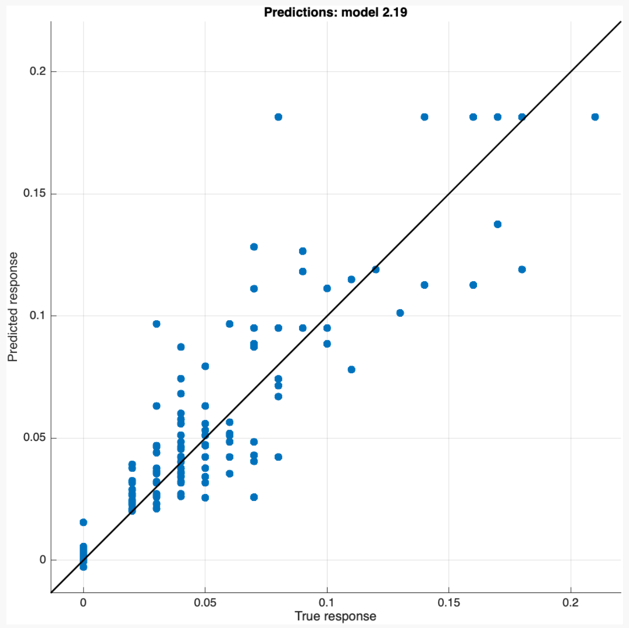
0.90



Ensemble of Trees
(Boosted Trees)

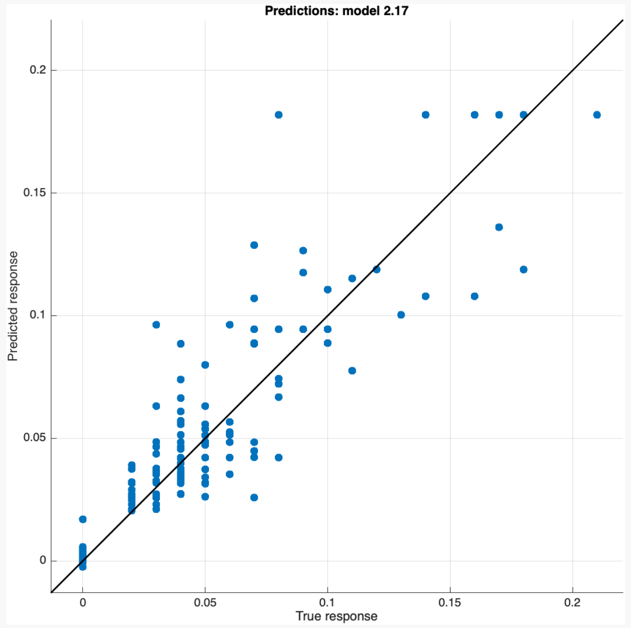
0.90

Residual solvent content



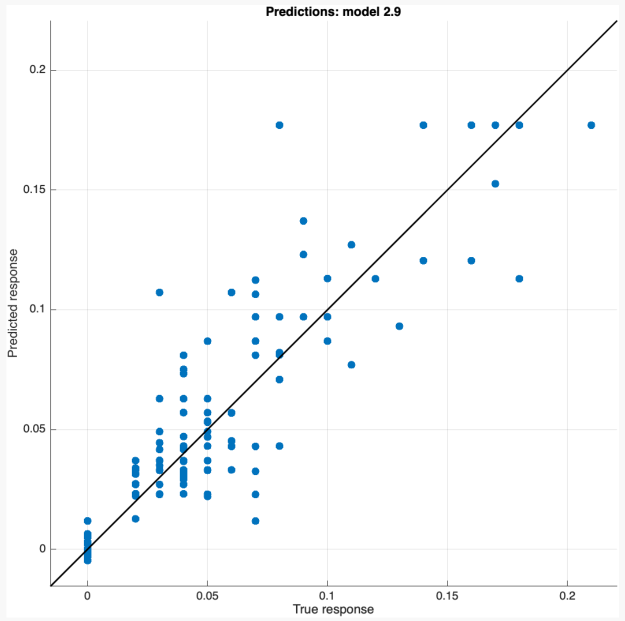
Gaussian Process Regression
(Rational Quadratic kernel)

0.82



Gaussian Process Regression
(Matern 5/2 kernel)

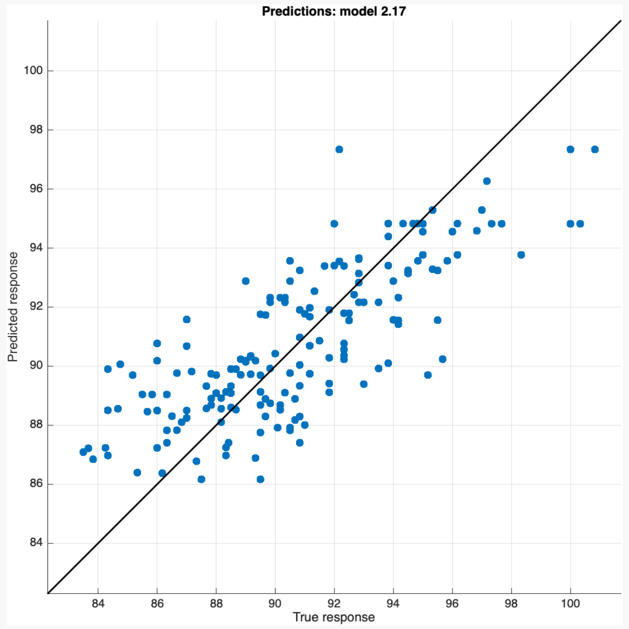
0.81



Support Vector Machine
(Quadratic kernel)

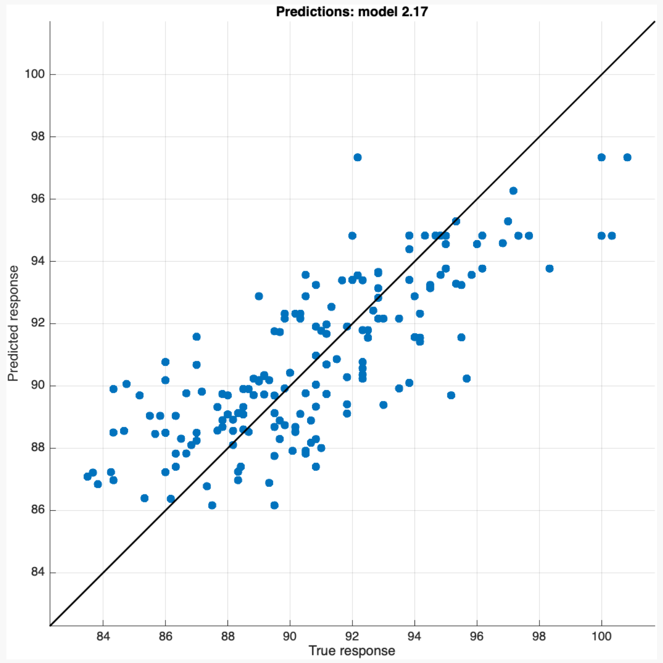
0.80

Percentage Dissolution



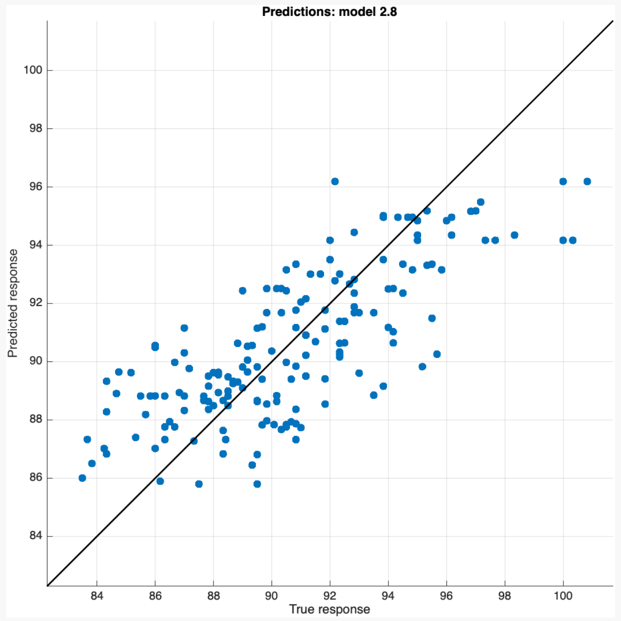
Gaussian Process Regression
(Squared exponential kernel)

0.60



Gaussian Process Regression
(Matern 5/2 kernel)

0.60



Support Vector Machine
(Linear kernel)

0.58

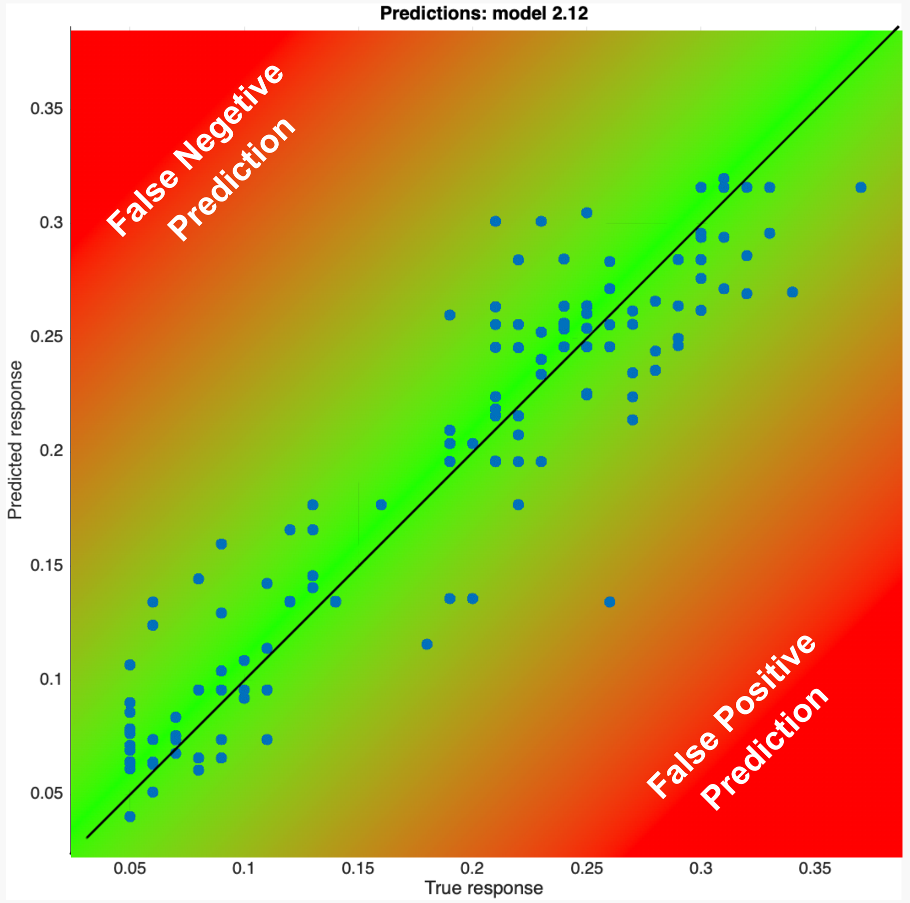
Legend

● Observations

— Perfect prediction

Limitations

- The R-squared score for Impurities is good and for Residual solvent is acceptable, but for Dissolution it's slightly low. However, this could improve with a bigger dataset.
- There were a few cases of False Positive Prediction (Eg: the batch actually had high impurity but the model predicted normal levels of impurity). This could lead to losses for the manufacturer, however, such cases are very low.
- There are chances of False Negative Prediction (Eg: the batch actually had normal impurity but the model predicted high levels of impurity). At least in this dataset, such cases were almost non-existent.



Total Percentage Impurity

Support Vector Machine
(Medium Gaussian kernel)

0.90

Future Scope

- More quality control data can be made public so that more robust models can be built for the prediction of efficacy.
- New features, that have a good correlation to final product efficacy, especially dissolution, can be discovered through scientific research.
- Regulatory authorities can maintain a central database of quality control data so that all manufacturers can benefit through shared data.

Flipkart

GRiD

5.0

Thank you