University of Padova

Department of Mathematics

Master Thesis in Data Science

Cationic Sulfur Trioxide Prediction in White Base Production

Supervisor
Prof. Alberto Roverato
University of Padova

Co-supervisor
Francesco Porpora
Procter and Gamble

Master Candidate
Sir. Sercan Albut

Student ID
2008494

Academic Year
2022-2023

“I suppose therefore that all things I see are illusions; I believe that nothing has ever existed of everything my lying memory tells me. I think I have no senses. I believe that body, shape, extension, motion, location are functions. What is there then that can be taken as true? Perhaps only this one thing, that nothing at all is certain.”
— Rene Descartes

Abstract

The production of high-quality white base, a crucial component in the paper, textile and cleaning products industries, is influenced by various parameters, including the concentration of cationic sulfur trioxide (simply Cat SO_3) which is a surfactant in the process. Elevated Cat SO_3 levels can adversely affect the product's properties, resulting in suboptimal quality and increased costs. This study proposes a data-driven approach to predict Cat SO_3 levels in white base production using data science and statistical data analysis techniques.

Traditional methods of Cat SO_3 level prediction often rely on manual monitoring and experience-based adjustments, meaning it can only be calculated after the white base production by a laboratory result, leading to inefficiencies, loss of valuable time and if it is not in the acceptable level, loss of the product . To address this, I employ advanced data science methodologies to develop an accurate predictive model. The proposed model integrates historical process data including laboratory results, and production parameters to forecast Cat SO_3 levels.

The methodology encompasses several key steps: data collection and pre-processing, explanatory data analysis (EDA), feature selection, model training, validation, and performance evaluation. Various machine learning algorithms, including regression techniques, are explored to identify the most suitable model for predicting Cat SO_3 levels. Process engineering techniques and engineers support are employed to extract relevant information from the complex and multivariate dataset.

The model's effectiveness is evaluated using real production data from Procter and Gamble HDL facility. Performance metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE),Root Avarage Square Error(RASE) and coefficient of determination (R^2) are employed to assess the accuracy of predictions. Additionally, the model's robustness and generalization capability are tested against unseen data to ensure its practical utility. Since laboratory analysis also tests more than one results of other value parameters, future Cat SO3 lab analysis results will be available for us to compare the results to new production data.

The results indicate that the proposed data-driven approach can replace traditional methods, yielding more accurate and consistent predictions of Cat SO_3 levels in white base production. This research contributes to the optimization of production processes, reducing wastage, and answers the value creation process which facility needs to complete for savings. The application of data science techniques not only explores Cat SO_3 level prediction but also establishes a foundation for further process optimization and automation in the white base manufacturing industry.

Contents

Listing of figures

Listing of tables

Listing of acronyms

FTC	Fundamental Theorem of Calculus
CatSO$_3$	Cationic Sulfur Trioxide
GLR	Generalized Linear Regression
GLR-L	Generalized Linear Regression Lasso Penalization
GLR-E	Generalized Linear Regression ElasticNet Penalization
RF	Random Forest
BF	Bootstrap Forest
SVM	Support Vector Machines
SVR	Support Vector Regressor
NN	Neural Network
kg/h	Killogram per Hour
N	Number of Observations
CLP	Continuous Liquid Proccessing
FIC	Flow Indicating Controller
PV	Present Value
SP	Set Point
OUT	Output
PID	Proportional Integral Derivative
BIS	Business Intelligence Service
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
R-Squared	Coefficient of Determination
RSS	Residuals Sum of Squares
TSS	Total Sum of Squares
AAE	Absolute Average Error
Freq	Frequency
ML	Machine Learning

1
Introduction

In early times, when the knowledge of nature was small, little attempt was made to divide science into parts, and men of science did not specialize. Aristotle was a master of all science known in his day, and wrote indifferently treatises on physics or animals. As increasing knowledge made it impossible for any one man to grasp all scientific subjects, lines of division were drawn for convenience of study and of teaching. Besides the broad distinction into physical and biological science, minute subdivisions arose, and, at a certain stage of development, much attention was, given to methods of classification, and much emphasis laid on the results, which were thought to have a significance beyond that of the mere convenience of mankind.

But we have reached the stage when the different streams of knowledge, followed by the different sciences, are coalescing, and the artificial barriers raised by calling those sciences by different names are breaking down. Geology uses the methods and data of physics, chemistry and biology; no one can say whether the science of radioactivity is to be classed as chemistry or physics, or whether sociology is properly grouped with biology or economics. Indeed, it is often just where this coalescence of two subjects occurs, when some connecting channel between them is opened suddenly, that the most striking advances in knowledge take place. The accumulated experience of one department of science, and the special methods which have been developed to deal with its problems, become suddenly available in the domain of another department, and many questions insoluble before may find answers in the new light cast upon them. Such considerations show us that science is in reality one, though we may agree to look on it now from one side and now from another as we approach it from the standpoint of physics, physiology or psychology.[1].

In today's rapidly changing world, the integration of different scientific disciplines has become vital in unlocking new insights and solving complex problems. One such collaboration that holds immense potential is the coalition between Data Science and Chemistry. This partnership brings together the analytical power of data science and the fundamental knowledge of chemistry, enriching both fields and contributing to groundbreaking advancements. By highlighting the importance and impact of this combined approach, we can understand how it is revolutionizing scientific research and opening doors to previously unexplored possibilities.

Data science has proven to be a game-changer in many scientific domains, and chemistry is no exception. With the ability to extract meaningful insights from vast amounts of data, data science plays a pivotal role in accelerating chemical research. There are many examples such as drug discovery, development, materials science and catalyst design which data science has been used to accelerate the researches. For example, researchers used machine learning algorithms to predict the outcomes of drug combinations for treating complex diseases. Their findings led to the discovery of a powerful new drug combination that showed extraordinary effectiveness against drug-resistant strains of the disease ending the "one drug one disease" era[2] Another example is that the researchers used data science and machine learning to develop an efficient catalyst for carbon dioxide (CO2) reduction. By leveraging computational models and big data analysis, they identified a new catalyst that significantly enhanced CO2 conversion efficiency, paving the way for sustainable energy solutions[3]The coalition between Data Science and Chemistry is transforming scientific research and pioneering groundbreaking discoveries. The integration of these two fields brings together the power of data analysis and domain expertise, resulting in advancements across various domains, including drug discovery, materials science, and process engineering which will be our main focus in this paper. By fostering collaborations and nurturing this synergy, we can continue to push the boundaries of scientific knowledge and address the most pressing challenges of our time.

In this paper our focus will be fixed on liquid white base production in Procter and Gamble HDL (Heavy Duty Laundry) facility located in Pomezia, Italy. The factory only produces HDL liquid and is considered a small plant. There the liquid production some of the ingredients are also produced here and some of the raw materials brought by suppliers. Production, packaging, storage and delivery, all happens in this plant. The coalition of a process engineering and data scientist will be explored and answer to the question of how each field help one another will be given. A process engineer in a chemical liquid production facility benefits significantly from collaborating with a data scientist for several reasons: Data-driven decision-making becomes possible, as data scientists possess the expertise to collect, analyze, and interpret large volumes of data from various sensors and sources. Leveraging advanced data analysis techniques, they provide actionable insights that allow process engineers to make informed decisions about process optimization and troubleshooting. Predictive models developed by data scientists can improve efficiency, reduce waste, and enhance product quality by continuously monitoring and adjusting processes. Early anomaly detection, root cause analysis, quality prediction, and resource optimization are among the many advantages, helping ensure regulatory compliance, reduce costs, foster innovation, and maintain data security. This partnership creates a culture of continuous improvement, allowing the facility to operate more efficiently and effectively in a competitive, ever-evolving industry. As you can guess by reading the title, our focus will be on the predictive modeling of a key ingredient which is called cationic sulfur trioxide in the white base production process.

Cationic sulfur trioxide (from here on we will call it CatSO3) is one of the dozen measures required to be checked in a detergent. The reason it is chosen to be analysed is that the value of it needs to be controlled and after production, white base's CatSO3 value measured by laboratory analysis. More than dozen ingredients are used to create white base. Since chemical reactions tend to have mathematical equation properties, the idea of soft end result measurement comes to the surface. Production of white base process is a closed process meaning no external effect can cause a reaction. Vacuumed tubes and mechanised process keeps person interaction impossible. Every ingredient and each process element is tracked by sensors. When the production process starts the product and all the ingredients are kept inside of pipes leaving no gap for external interaction until it is poured in to the bottles. Specifically for producing detergent, white base is just the beginning. After producing white base production continues to a different stage to produce the detergent. Since we are interested in white base production cycle, we will keep other elements and production elements out of the scope. For example some ingredients used to produce white base also processed before reaching their dosing point into the system. To make sure that our point of view is limited and clear we won't take into account of previous and after math process of materials. It is important to mention that before manufacturing phase has no ongoing chemical reactions by the time they reach to the white base production start. So that there won't be any extra work to be included in our analysis mentioning the prologue.

The production of liquid soap is made by mixing, in a blending process, different raw materials. The most abundant one is the white base (alongside with water), that itself is a mixture of 10/13 materials. To be used for production scopes, the white base must follow the manufacturing standard requirement. In the process there are more than one scope needs to be check to be able to abide manufacturing standards. Two of the most important scopes are CatSO3 value and ph value of the white base. Even though there is a pH probe online, the white base is a very viscous liquid that does not allow to rely on this measure to assess its quality. Reason why, for each production, every 30-45 minutes an operator from the control room has to go to the line, take a sample of the produced white base and send it to the laboratory. Here the sample is diluted to 10 percent of its concentration, making the initial solution an aqueous solution for which the probe can give a more precise result. The result from the analysis of the pH and CatSO3 is then sent back to the control room operator that reading the results and knowing the pH set point of the white base, trims (increases or decreases for a small percentage) the flow set point of one of the raw materials (in ph case the caustic soda since it is the most impacting on pH). In the process control room we have the option to change the setpoints. Depending on the formula card in production, the setpoint of each raw material is dictated by the manufacturing standard. This one gives us the setpoint and the limit within which the production can continue. The only setpoint that can be changed is the Caustic Soda setpoint. In the panel, the operator of the control room can change the setpoint of its flowrate (to compensate the flowrate changes, of course the amount added/subtracted from the Caustic soda flowrate is inversely subtracted/added from the water flowrate) just inputting the percent of how much they want to increase/decrease.

Current studies in research and development center showed us that ph value can be controlled by soft sensing. Using the power of data we can control the ph of the product while it is still in production. After this process integrated to the current work process, the company wants to eliminate laboratory analysis completely. To be able to do that they need to know the value of CatSO3 as well and that brings us the use case of this paper. What we want to achieve is to predict CatSO3 level in the whitebase while it is still in production. Doing so will help company to eliminate the laboratory analysis completely. If we can model the level of CatSO3 by the ingredients of the process, we will achieve our goal. In the case of CatSO3 there is no controllable variable in the process. However with enough historical laboratory results and time series data of ingredients we might able to achieve prediction with low error, so to say enough to be acceptable. By achieving so this project will help the company have soft gains, meaning the time spent on this work process will be regained and will be directed elsewhere.

To be able to understand the process we will share with you insights from process and chemistry engineers' perspective. Without the explanation and process knowledge, data is just pile of values. Guidance goes both ways. While the process engineers help us to understand the process by doing so direct us how to use the data accordingly, we will help them to identify and shape patterns which are not easy to be seen by naked eyes. As complexity increases, the cooperation should increase as well to be able to come up with usable answers. As we continue explaining the process some details which are not crucial for our topic will be generalized, over simplified

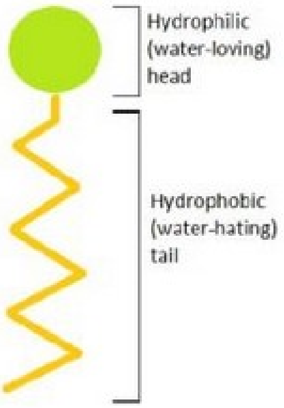


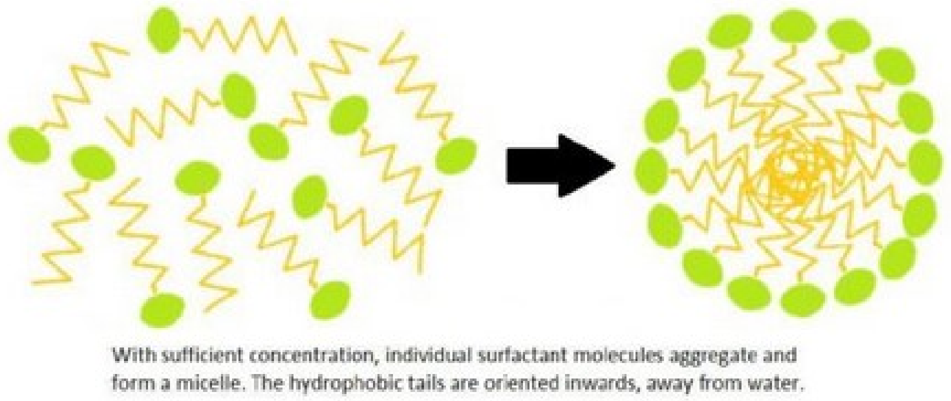
Figure 1.1: Surfactants

and names will be changed so that process will keep its uniqueness and secrets as the corporate policy dictates non disclosure. The information that we share here is also publicly available.

Detergent production involves several key steps in the process engineering aspect, from raw material selection to final product packaging. Raw material selection and production part also includes two aspect. Preparation of white base then preparation of detergent. White base is the actual cleaning product which is made of surfactants and builders. Surfactants are often acidic in nature. To make them suitable for use in detergents, they undergo neutralization with alkaline substances such as sodium hydroxide (caustic soda). This step is crucial for achieving the desired pH of the detergent. There are other builders to make whitebase stable and ofcourse water is one of the main ingredients. Other builders include sodium in them. From the process point of view, to better understand what we are looking for, we should know what is surfactant in the first place.

Surfactants are a primary component of cleaning detergents. The word surfactant means surface active agent. As the name implies, surfactants stir up activity on the surface you are cleaning to help trap dirt and remove it from the surface.Figure 1.1 Surfactants have a hydrophobic (water-hating) tail and a hydrophilic (water-loving) head. The hydrophobic tail of each surfactant surrounds soils. The hydrophilic head is surrounded by water. When there are a sufficient amount of surfactant molecules present in a solution they combine together to form structures called micelles. As the micelle forms, the surfactant heads position themselves so they are exposed to water, while the tails are grouped together in the center of the structure protected from water.The micelles work as a unit to remove soils. The hydrophobic tails are attracted to soils and surround them, while the hydrophilic heads pull the surrounded soils off the surface and into the cleaning solution. Then the micelles reform with the tails suspending the soil in the center of the structure. Figure 1.2 Cationic surfactants have a positive charge on their hydrophilic end. The positive charge makes them useful in anti-static products, like fabric softeners. Cationic surfactants can also serve as antimicrobial agents, so they are often used in disinfectants. Figure 1.3 [4]

The process follows precisely placed pipelines and dosing points. While going through the pipes the solution is continuously mixed. There are filler tanks which the solution mixed again then put through the pipes again to make sure the solution is homogeneous. Every dosing valve has a flow meter and pressure sensor to check the stability of the process. Temperature and pressure is constantly checked. After the full composure is reached the



With sufficient concentration, individual surfactant molecules aggregate and form a micelle. The hydrophobic tails are oriented inwards, away from water.

Figure 1.2: Micelles

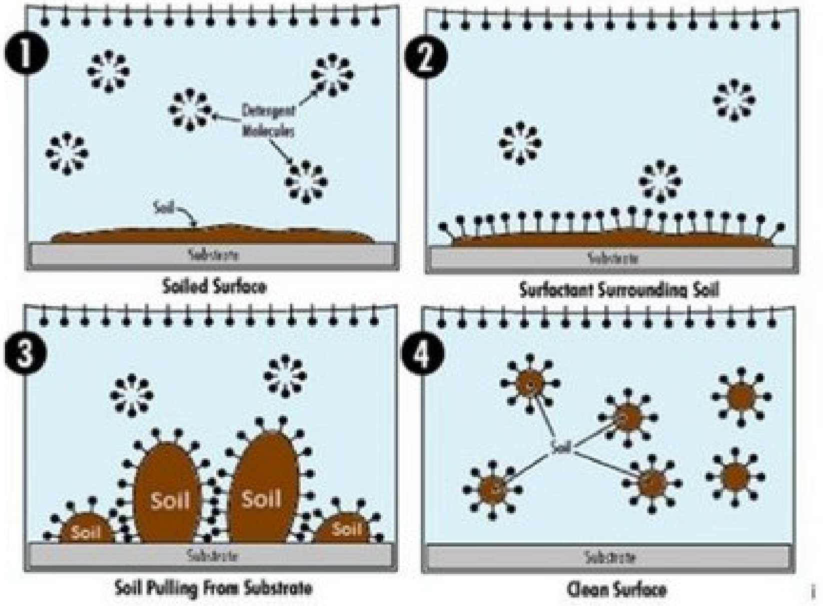


Figure 1.3: How cleaning works

whitebase is stored in the tanks. Right before tanks the solution passes through a thin pipe which allows operators to take samples from the final product.This sample is used for recording the detailed laboratory analysis of the final whitebase.

Laboratory is located next to the production site. The reason is that the product has highly active formula and time is the essence of the results. That is why sampling and analysis happens simultaneously. Laboratory analysis consists of multiple results including ph, viscosity and $CatSO3$ results. There are certain parameters which the mixture should be in between on the account of multiple variables. Laboratory technicians starts sampling right after the production reaches a stable condition and through the production in certain time lapses it continues. It also helps process controllers to adjust ingredients according to results. For quality assurance the lab analysis are our corner stone of the production. Regular and systematic lab analysis, combined with real-time monitoring and feedback systems, forms the basis of effective process control in detergent production. It allows for the early detection of deviations from the desired specifications, enabling prompt corrective actions to maintain product quality and consistency.

Process engineering and data science knowledge collaboration lead us to believe that through the process some of the laboratory results could be predicted by using historical dataset of the production ingredients. In this project the aim is to predict $CatSO3$ level of the whitebase through historical data (time series data)of ingredients. Since every lab sample deliver one result and sampling happens certain time lapses the study won't be a time series analysis but lab results will be time stamped with the exact time as the samples have been taken. So that later on time stamps will be matched with the exact times of ingredient's present values. That means if the sample has been taken at ten o'clock, we will only take the data of ingredients' values at ten o'clock.

The process will include data cleaning, regression analysis, dependent variable analysis, explanatory data analysis and comparison of predictive modelling but there are few points we should be considering before diving into data science of the process. First of all this is a chemical process happening in real life. Even though people consider chemical processes as two plus two equals four on paper, in real life production there are deviations caused by natural process itself. What we are considering is that data coming out of the sensors will be our pin point. Due to high volume of production, viscosity of the fluids, activation levels of ingredients, natural state of production involvement, lab analyst human error and sensor accountability the delivered results will have a certain deviation from the exact real supposed values.

In the upcoming data set section the data set will be explained according ingredients and response result. Ingredients will have code names not to disclose full formulas of the production but there will be small explanations of each of them. The topics of the thesis are delivered as follows: Ingredient disclosure and description, cleaning, pre-processing, feature engineering. Later on The Generalized Linear Regressions, Adaptive Lasso and Adaptive Elasticnet (GLRL and GLRE), Bootstrap Forest(RF), Support Vector Machine (SVM) and Neural Network (NN) algorithms are explained in the chapter models and the methods of the models and their evaluation applied in the research. Finally, chapter 4 will include the performance evaluation of all the models and evaluation criteria of the models with higher specifications.

2
Dataset

The production of liquid soap is made by mixing, in a blending process, different raw material. The most abundant one is the white base (alongside with water), that itself is a mixture of 10/13 materials. To be used for production scopes, the white base must follow the manufacturing standard requirement. This one gives to the department the guideline, setpoint and range limits that have to be taken, for example pH, Viscosity, CatSO3 and appearance. In this Dataset we will focus on CatSO3 levels. We will see and discuss how it has been tried to predict the CatSO3 of a mixture basing it on historical production data.

In Pomezia plant the production line is represented by Figure 2.1 below.

The pipeline has deflector inside. At specific meters are present the raw material injection point that mix and reach the storage tank. At the end of the line there are two probes: the first one is a temperature probe and the last one is a pH probe. Right before temperature probe we have our sampling valve. The sampling happens depending on the length of the production. These are 4 to 6 hours productions depending on the need. Each production is called "Run". Every Run has at least 4 sampling events, the number might increase to 6. Shortest runs have 4 sampling as first sampling happens after production starts and becomes stable. Second happens in the middle of production, third happens right before finishing the production and last sampling happens for Tank which we will exclude since it is not taken while the production runs. The Runs are uninterrupted from the start and finish. Stable state means every ingredient reaches their set points and will be on that point until production stops. So we can assume that flows are stable but includes variation. According to production values sometimes set points are changed while production runs. According to temperature, ph and sampling results of the product while it is running, adjustments are made to keep the product's CatSO3, ph, apperance and viscocity values in check. that is why sampling happens in between.

The aim of the project as previously stated, using the historical ingredients data to predict the CatSO3 laboratory results, using predictive modelling methods. After the creation of model the plant will be able to reduce the sampling to 1 which will be taken from the Storage Tank to finalize the process.

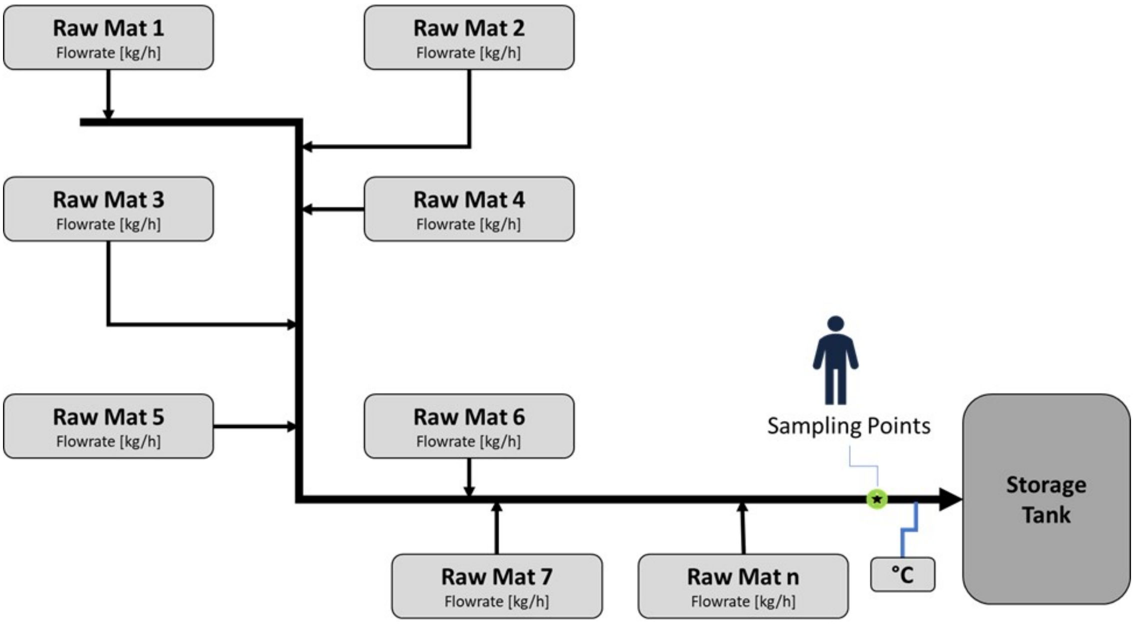


Figure 2.1: Process of Whitebase Production

2.1 Data Shape

Due to nondisclosure agreement the ingredients will have coded shorten names. This information is strictly protected since it reveals the product combination and could be replicated via out sources. Ingredient table Table 2.1, shows us that there are many ingredients to this process but it is important to identify which ones are important for our result. Sampling results are scarce. This data is collected between March 2022 and January 2023. There are only 526 acceptable sampling results for our continues ingredient data set. So our sample's N value is 526. Except Brand (Categorical), all the values are numerical floats. We are receiving ingredients data as time series but since we are exactly matching them with the time of laboratory results, time is irrelevant for the models.

Brand is a diverse topic for the study. While it might be useful to the model we can not accept it as a variable. Reason is that brand changes periodically and once a brand decided as terminated by general office, it won't be re introduced to the production. In the upcoming model creations brand will be discarded so categorical values will be eliminated. How ever Brand will help us to create validation samples in the upcoming sections, so that every validation will be grouped by significant brand. That will help us increase the model accuracy while we validate and test the model accordingly. There are other component highly correlated to brand which will be included in the model, so brand will not be included in model creation but it will make validation-test sampling very easy.

2.2 Data Cleaning and Variable Selection

While the process is going sometimes sensors that reads the dosing values give unacceptable readings. That is why we would like to clean dataset first and see the outliars. To do this we have created the time series results of

Code Name	**Specification**
PHIC401	Line pH
FIC104	Material Flowrate (kg/h)
FIC108	Fatty Acid Flowrate (kg/h)
FIC155	Material Flowrate (kg/h)
FIC156	Material Flowrate (kg/h)
FIC158	Material Flowrate (kg/h)
FIC181	Material Flowrate (kg/h)
FIC310	Material Flowrate (kg/h)
FIC2-131	Wash Recipe
pHCLP	Sample pH
FIC102	Material Flowrate (kg/h)
FIC105	Surfactant 1 Flowrate (kg/h)
FIC107	Material Flowrate (kg/h)
FIC112	Material Flowrate (kg/h)
FIC401	Water Flowrate (kg/h)
FIC175F-L	Surfactant 2 Flowrate (kg/h)
FIC8-401	Material Flowrate (kg/h)
FIC1-115	Reblend 1 Flowrate (kg/h)
FIC2-115	Reblend 2 Flowrate (kg/h)
FIC3-115	Reblend 3 Flowrate (kg/h)
Brand	Brand of production
CatSO3	CatSO3 level of sample

Table 2.1: Ingerdients

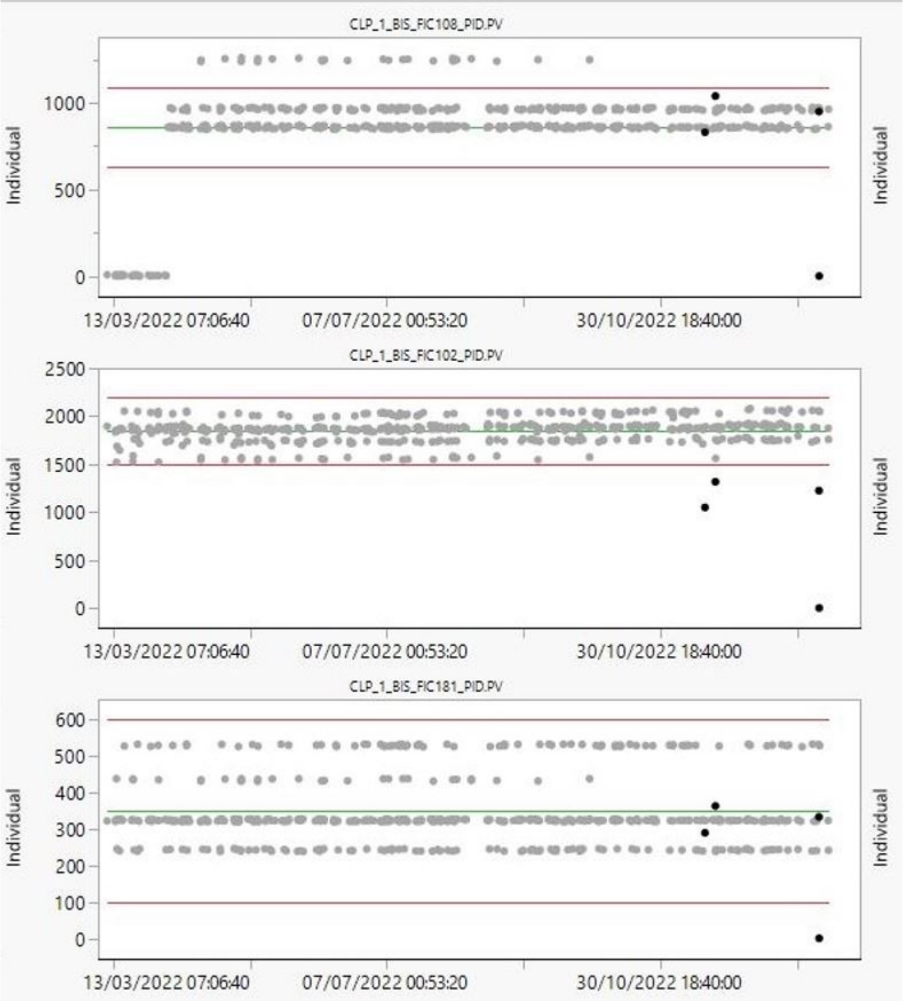


Figure 2.2: Outliars 1

each ingredient. Rather than share all of the ingredients results I will simply share the ingredients which shows obvious outliars. Below 3 figures shows us main graphs that indicate outliars. Figure 2.2 shows the outliars on the ingredient FIC108, FIC102 and FIC181. The highlighted points in each graph corresponds to the same row of information in the dataset. Figure 2.3, Figure 2.4 shows the same points being outliars for ingredient FIC156 and FIC105. We proceed to remove the outliars. Finally we check the ph value time series, both sample and inline Figure 2.5. Even after looking at the values, from a point of process engineering and a data scientist highlighted values are showing out of scope values. After identifying all the outliars and removing them our sample size is reduced to 519.

Also lets see the distribution of our response variable and ph values of the whole dataset Figure 2.6. As you can see CatSO3 values are grouped between several points while ph seems normally distributed. As we previously mentioned Brand plays a specific role in the production and brands have different values of CatSo3. Case of reblends are insignificant since it means reblending the product into same production brand. As you can see

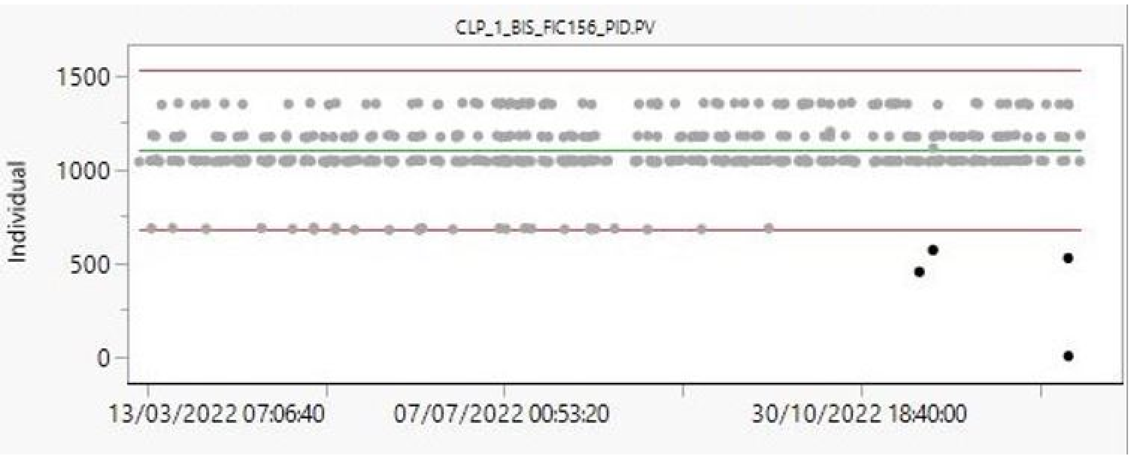


Figure 2.3: Outliars 2

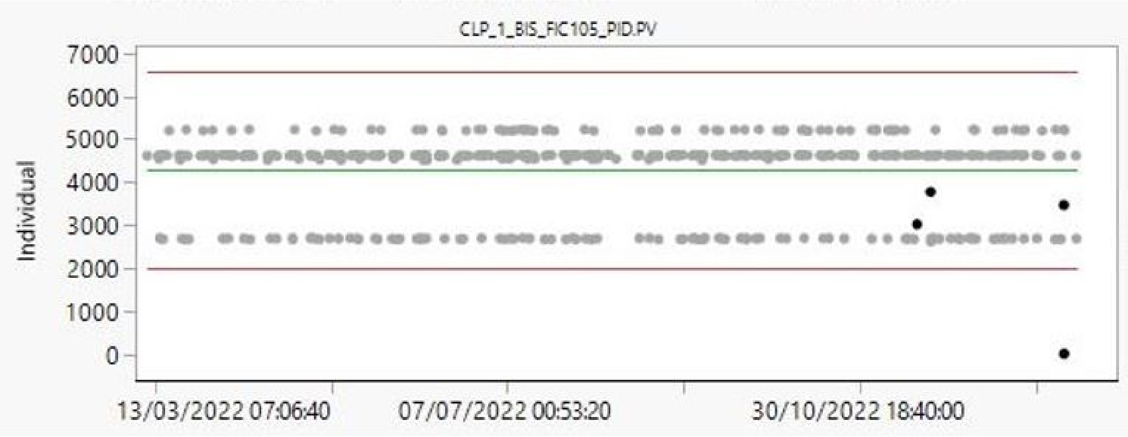


Figure 2.4: Outliars 3

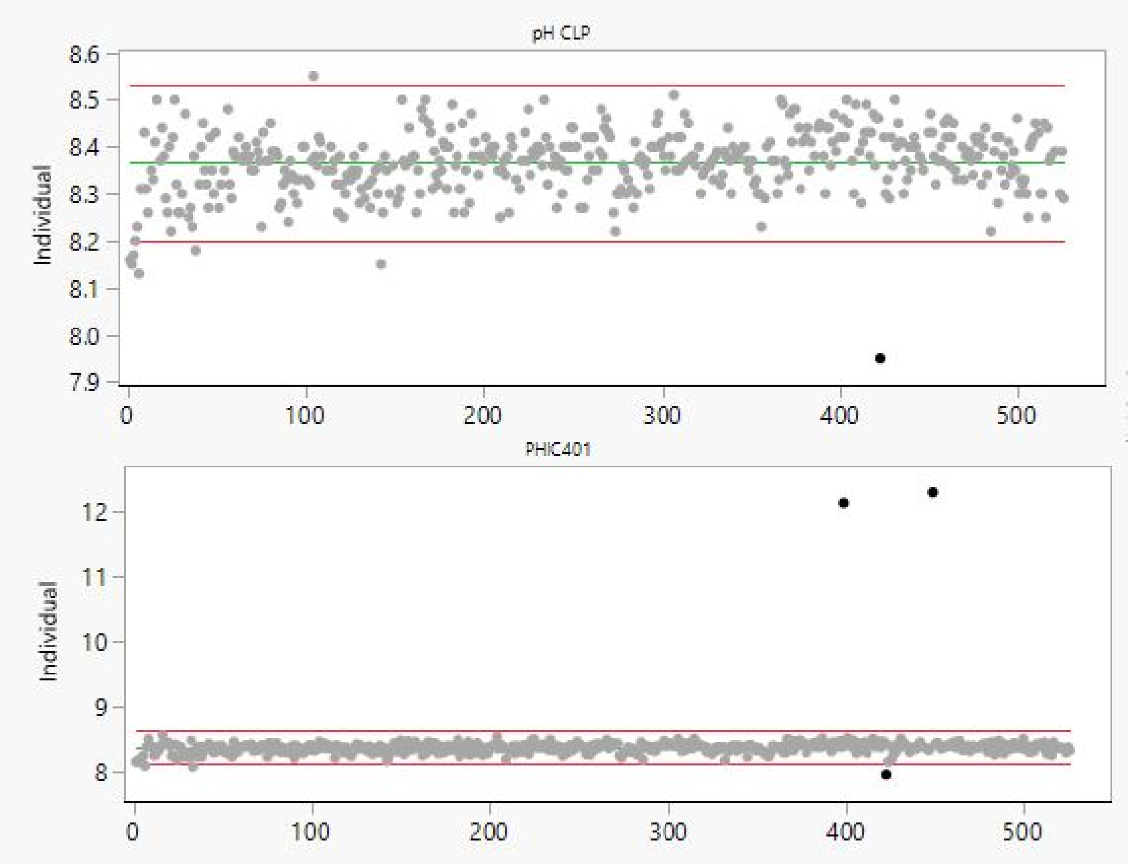
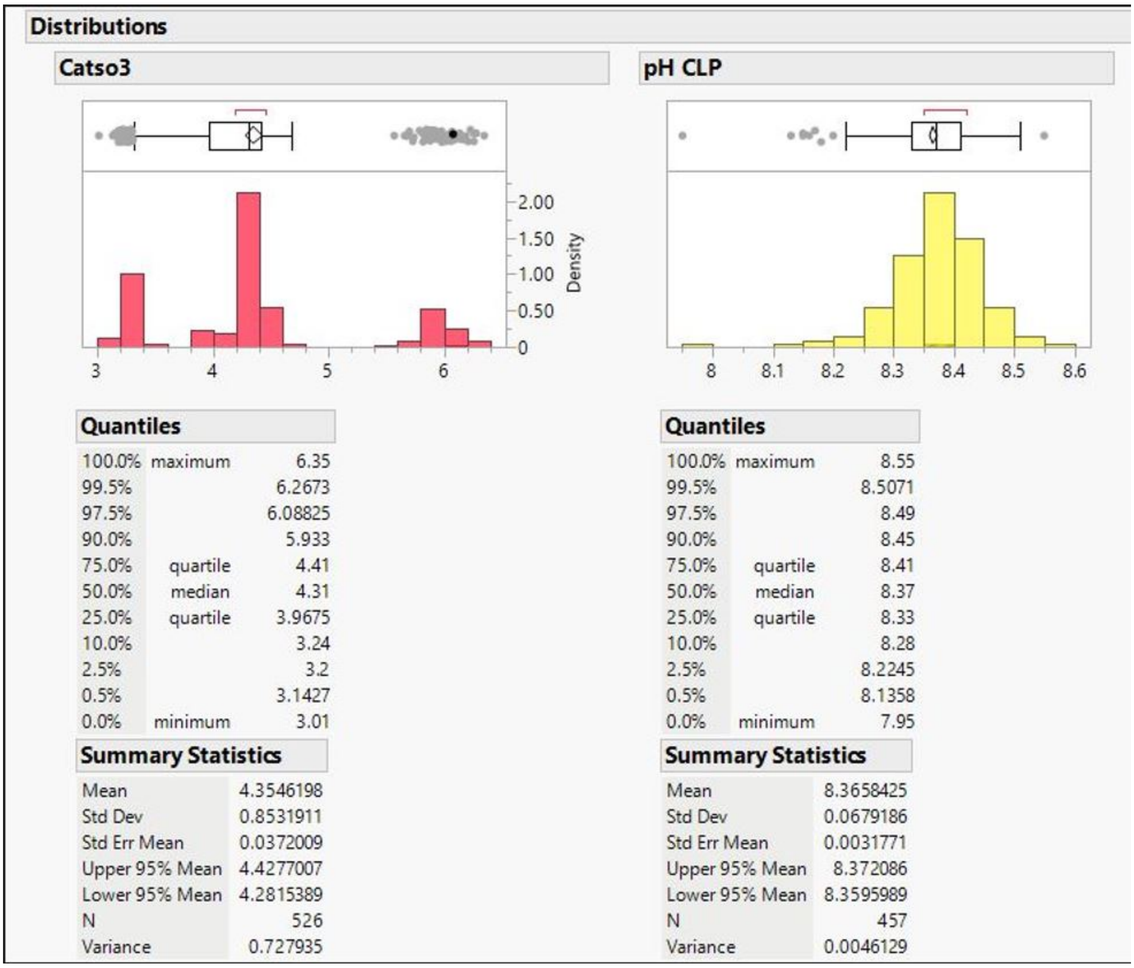


Figure 2.5: Outliars ph



Catso3

Quantiles

100.0%	maximum	6.35
99.5%		6.2673
97.5%		6.08825
90.0%		5.933
75.0%	quartile	4.41
50.0%	median	4.31
25.0%	quartile	3.9675
10.0%		3.24
2.5%		3.2
0.5%		3.1427
0.0%	minimum	3.01

Summary Statistics

Mean	4.3546198
Std Dev	0.8531911
Std Err Mean	0.0372009
Upper 95% Mean	4.4277007
Lower 95% Mean	4.2815389
N	526
Variance	0.727935

pH CLP

Quantiles

100.0%	maximum	8.55
99.5%		8.5071
97.5%		8.49
90.0%		8.45
75.0%	quartile	8.41
50.0%	median	8.37
25.0%	quartile	8.33
10.0%		8.28
2.5%		8.2245
0.5%		8.1358
0.0%	minimum	7.95

Summary Statistics

Mean	8.3658425
Std Dev	0.0679186
Std Err Mean	0.0031771
Upper 95% Mean	8.372086
Lower 95% Mean	8.3595989
N	457
Variance	0.0046129

Figure 2.6: CatSO3 and Ph distribution

in the Figure 2.7 and Figure 2.8 how the CatSO3 levels are grouped. Reblend could be anyvalue which doesnt show significance. The problem is that we can not use brand categories as an input since they might go out of production in the future. There are currently four brands in the production: Ambiorix, Tigress, Arctica and Jupiler. However we have some ingredients where we always use in the production and they resemble similar difference and does the grouping as brand does. One of them is called surfactant and by chemistry knowledge that has been shared, they play really important role on describing other surfactants. The Figure 2.9 shows the graph of hand picked ingredients that resembles the grouping of brand. Also one of them is surfactant which means it plays big role in prediction of CatSO3.

The data includes many independent variable to be considered in the modelling part. Correlation maps will help us to further eliminate variables. For significance of their separation and importance the surfactant variables will be kept for modelling. The rest of the variables chosen be decided by using correlation matrix and fitting Multivariate Least Square Regression model backward stepwise selection. Figure 2.10 Shows us that between variables there are a lot of correlations. So decision started with surfactants and eliminating their correlated counterparts.

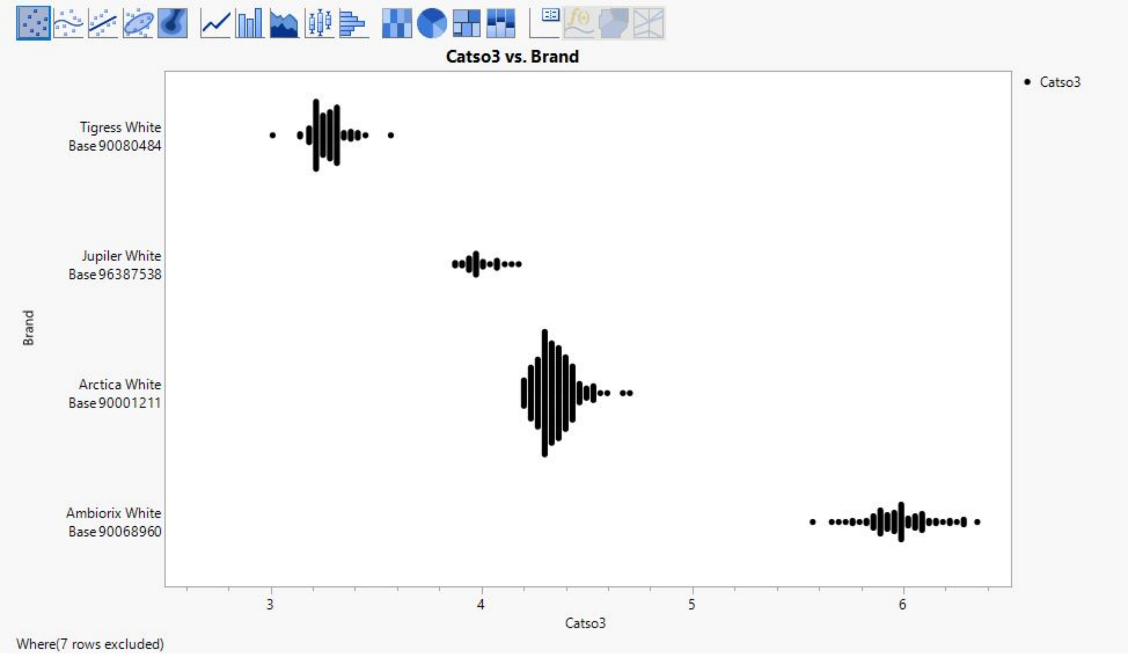


Figure 2.7: CatSO3 vs Brands

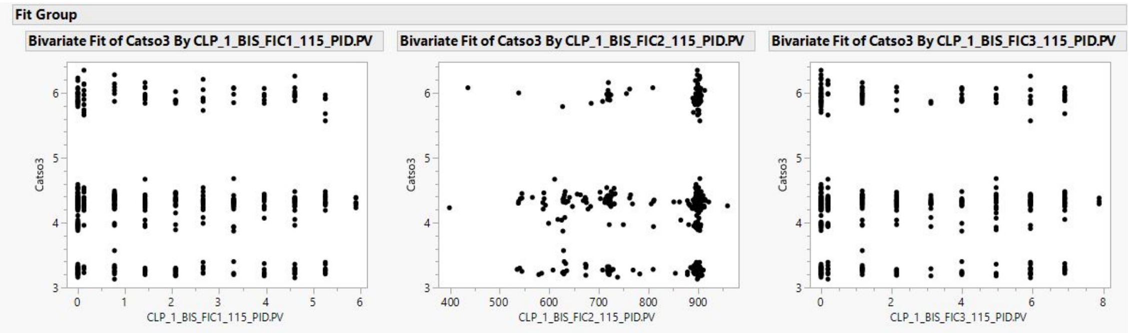


Figure 2.8: CatSO3 vs Reblends distribution

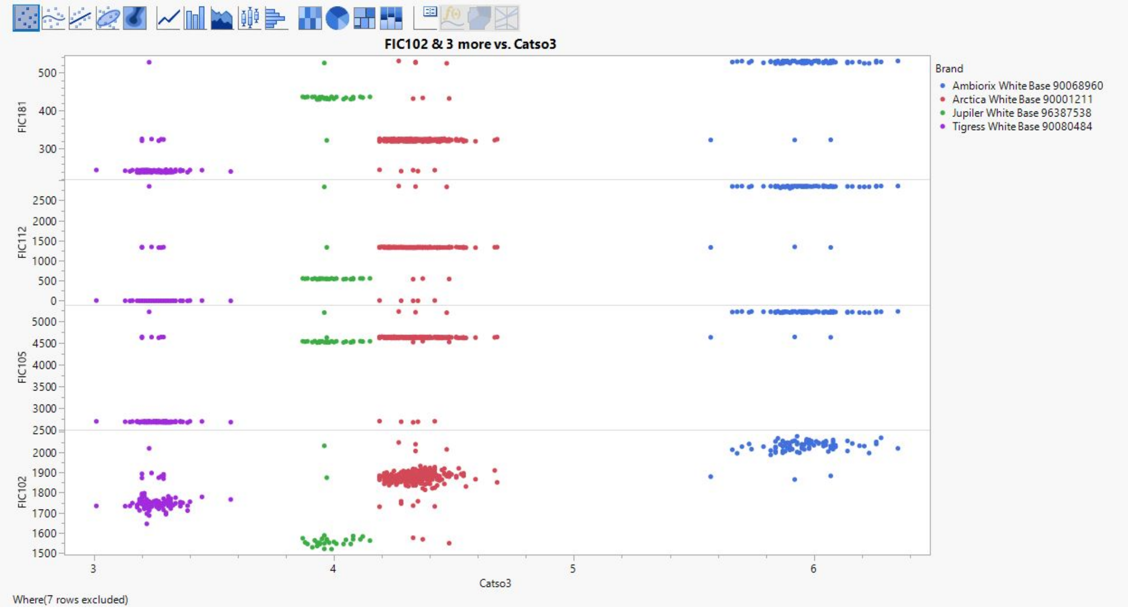
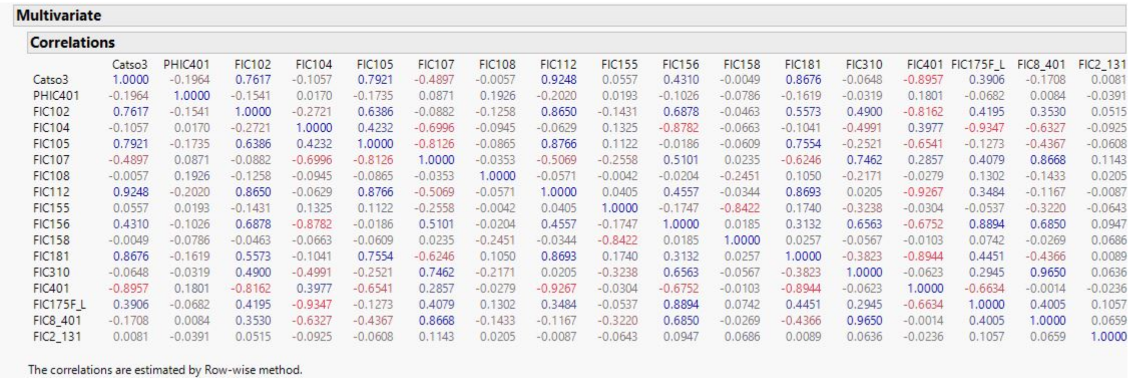


Figure 2.9: CatSO3 vs FIC102,FIC105,FIC112,FIC181

Later on most correlated variables have been chosen and their correlated counterparts have been removed. Also at the same time Multivariate Least Square Regression model backward stepwise selection applied simultaneously to see variable significance while they have been removed. Figure 2.11 Shows that there are insignificant variables included in our model. Though model is valid according to predicted results, by backward stepwise selection we reduced the model variables to 8. Figure 2.12 shows us the same results and with reduced residuals for predicted values. In the effect test part ph values have been shown in effective to the model via F test. However when we remove the ph from the model, results worsens as you can see in Figure 2.13. So we decided to keep ph value in the model. We come to this conclusion via looking R square and p values of the model and ingredients separately. Also prediction vs real graph proves our point. Feature selection is primarily focused on removing non-informative or redundant predictors from the model. Many models, especially those based on regression slopes and intercepts, will estimate parameters for every term in the model. Because of this, the presence of non-informative variables can add uncertainty to the predictions and reduce the overall effectiveness of the model.[5]. Certain statistical literature suggest us to use Principle component analysis to reduce the feature dimension but as process engineering part we need to consider features individually to control them. Combining features would not work for our solution.



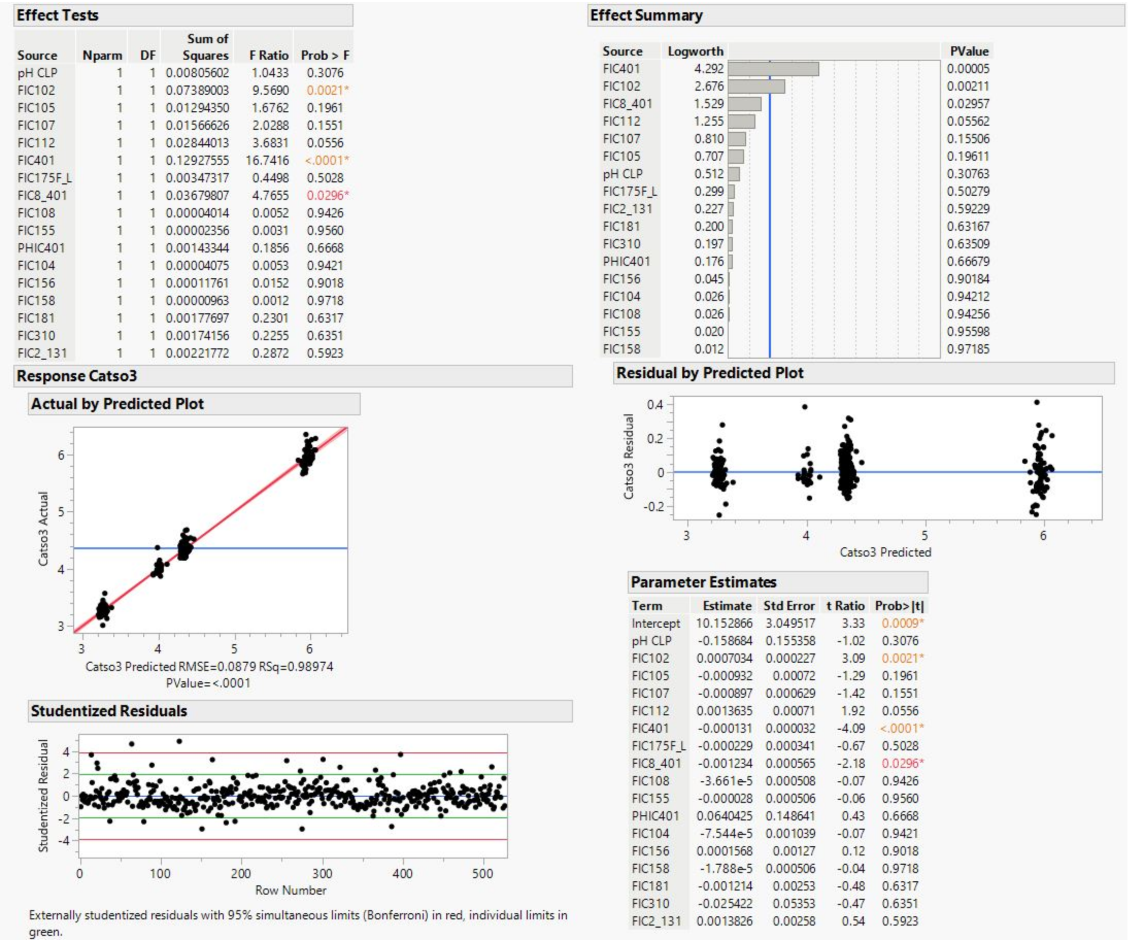
Multivariate

Correlations

	Catso3	PHIC401	FIC102	FIC104	FIC105	FIC107	FIC108	FIC112	FIC155	FIC156	FIC158	FIC181	FIC310	FIC401	FIC175F_L	FIC8_401	FIC2_131
Catso3	1.0000	-0.1964	0.7617	-0.1057	0.7921	-0.4897	-0.0057	0.9248	0.0557	0.4310	-0.0049	0.8676	-0.0648	-0.8957	0.3906	-0.1708	0.0081
PHIC401	-0.1964	1.0000	-0.1541	0.0170	-0.1735	0.0871	0.1926	-0.2020	0.0193	-0.1026	-0.0786	-0.1619	-0.0319	0.1801	-0.0682	0.0084	-0.0391
FIC102	0.7617	-0.1541	1.0000	-0.2721	0.6386	-0.0882	-0.1258	0.8650	-0.1431	0.6878	-0.0463	0.5573	0.4900	-0.8162	0.4195	0.3530	0.0515
FIC104	-0.1057	0.0170	-0.2721	1.0000	0.4232	-0.6996	-0.0945	-0.0629	0.1325	-0.8782	-0.0663	-0.1041	-0.4991	0.3977	-0.9347	-0.6327	-0.0925
FIC105	0.7921	-0.1735	0.6386	0.4232	1.0000	-0.8126	-0.0865	0.8766	0.1122	-0.0186	-0.0609	0.7554	-0.2521	-0.6541	-0.1273	-0.4367	-0.0608
FIC107	-0.4897	0.0871	-0.0882	-0.6996	-0.8126	1.0000	-0.0353	-0.5069	-0.2558	0.5101	0.0235	-0.6246	0.7462	0.2857	0.4079	0.8668	0.1143
FIC108	-0.0057	0.1926	-0.1258	-0.0945	-0.0865	-0.0353	1.0000	-0.0571	-0.0042	-0.0204	-0.2451	0.1050	-0.2171	-0.0279	0.1302	-0.1433	0.0205
FIC112	0.9248	-0.2020	0.8650	-0.0629	0.8766	-0.5069	-0.0571	1.0000	0.0405	0.4557	-0.0344	0.8693	0.0205	-0.9267	0.3484	-0.1167	-0.0087
FIC155	0.0557	0.0193	-0.1431	0.1325	0.1122	-0.2558	-0.0042	0.0405	1.0000	-0.1747	-0.8422	0.1740	-0.3238	-0.0304	-0.0537	-0.3220	-0.0643
FIC156	0.4310	-0.1026	0.6878	-0.8782	-0.0186	0.5101	-0.0204	0.4557	-0.1747	1.0000	0.0185	0.3132	0.6563	-0.6752	0.8894	0.6850	0.0947
FIC158	-0.0049	-0.0786	-0.0463	-0.0663	-0.0609	0.0235	-0.2451	-0.0344	-0.8422	0.0185	1.0000	0.0257	-0.0567	-0.0103	0.0742	-0.0269	0.0686
FIC181	0.8676	-0.1619	0.5573	-0.1041	0.7554	-0.6246	0.1050	0.8693	0.1740	0.3132	0.0257	1.0000	-0.3823	-0.8944	0.4451	-0.4366	0.0089
FIC310	-0.0648	-0.0319	0.4900	-0.4991	-0.2521	0.7462	-0.2171	0.0205	-0.3238	0.6563	-0.0567	-0.3823	1.0000	-0.0623	0.2945	0.9650	0.0636
FIC401	-0.8957	0.1801	-0.8162	0.3977	-0.6541	0.2857	-0.0279	-0.9267	-0.0304	-0.6752	-0.0103	-0.8944	-0.0623	1.0000	-0.6634	-0.0014	-0.0236
FIC175F_L	0.3906	-0.0682	0.4195	-0.9347	-0.1273	0.4079	0.1302	0.3484	-0.0537	0.8894	0.0742	0.4451	0.2945	-0.6634	1.0000	0.4005	0.1057
FIC8_401	-0.1708	0.0084	0.3530	-0.6327	-0.4367	0.8668	-0.1433	-0.1167	-0.3220	0.6850	-0.0269	-0.4366	0.9650	-0.0014	0.4005	1.0000	0.0659
FIC2_131	0.0081	-0.0391	0.0515	-0.0925	-0.0608	0.1143	0.0205	-0.0087	-0.0643	0.0947	0.0686	0.0089	0.0636	-0.0236	0.1057	0.0659	1.0000

The correlations are estimated by Row-wise method.

Figure 2.10: Corraletion matrix



Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
pH CLP	1	1	0.00805602	1.0433	0.3076
FIC102	1	1	0.07389003	9.5690	0.0021*
FIC105	1	1	0.01294350	1.6762	0.1961
FIC107	1	1	0.01566626	2.0288	0.1551
FIC112	1	1	0.02844013	3.6831	0.0556
FIC401	1	1	0.12927555	16.7416	<.0001*
FIC175F_L	1	1	0.00347317	0.4498	0.5028
FIC8_401	1	1	0.03679807	4.7655	0.0296*
FIC108	1	1	0.00004014	0.0052	0.9426
FIC155	1	1	0.00002356	0.0031	0.9560
PHIC401	1	1	0.00143344	0.1856	0.6668
FIC104	1	1	0.00004075	0.0053	0.9421
FIC156	1	1	0.00011761	0.0152	0.9018
FIC158	1	1	0.00000963	0.0012	0.9718
FIC181	1	1	0.00177697	0.2301	0.6317
FIC310	1	1	0.00174156	0.2255	0.6351
FIC2_131	1	1	0.00221772	0.2872	0.5923

Effect Summary

Source	Logworth	PValue
FIC401	4.292	0.00005
FIC102	2.676	0.00211
FIC8_401	1.529	0.02957
FIC112	1.255	0.05562
FIC107	0.810	0.15506
FIC105	0.707	0.19611
pH CLP	0.512	0.30763
FIC175F_L	0.299	0.50279
FIC2_131	0.227	0.59229
FIC181	0.200	0.63167
FIC310	0.197	0.63509
PHIC401	0.176	0.66679
FIC156	0.045	0.90184
FIC104	0.026	0.94212
FIC108	0.026	0.94256
FIC155	0.020	0.95598
FIC158	0.012	0.97185

Response Catso3

Actual by Predicted Plot

Catso3 Predicted RMSE=0.0879 RSq=0.98974
PValue=<.0001

Residual by Predicted Plot

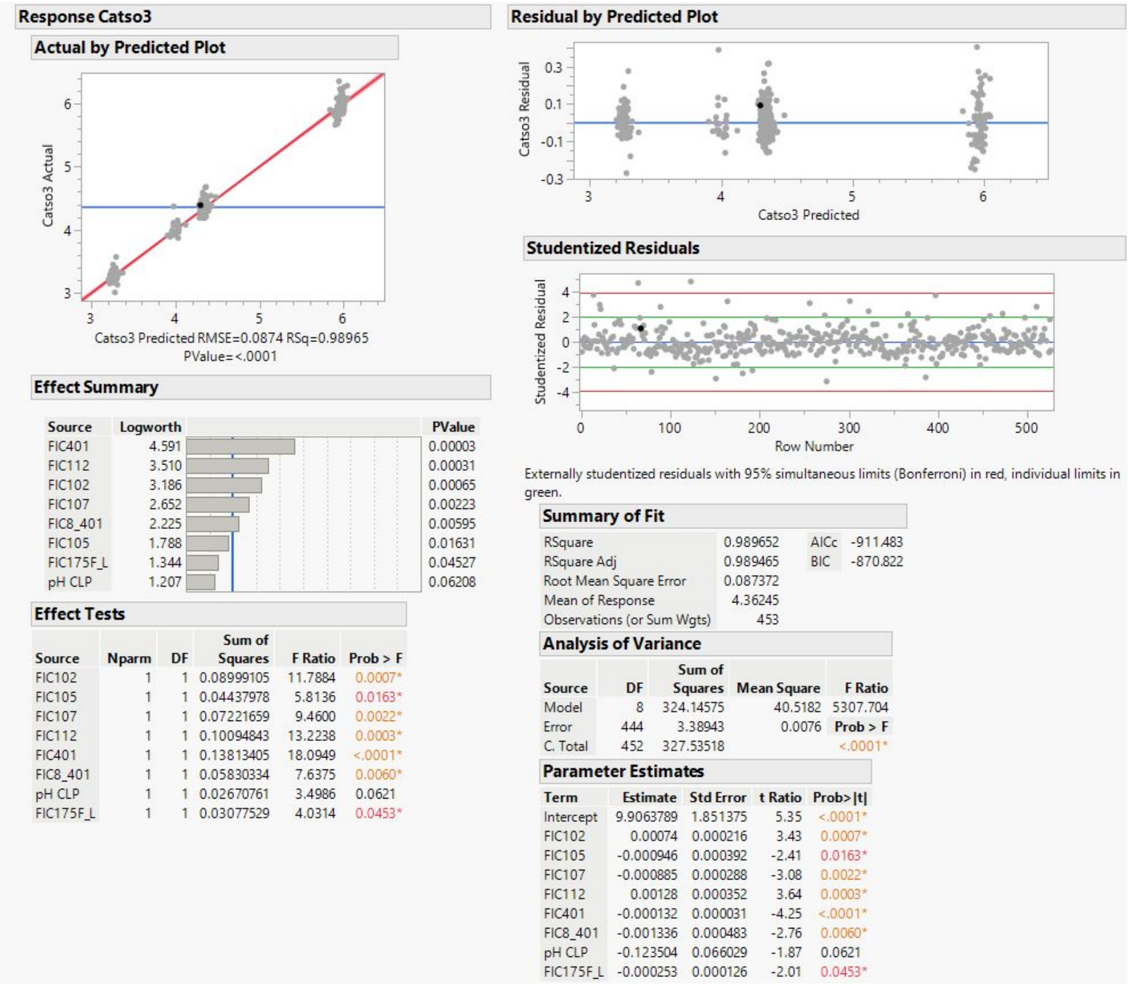
Studentized Residuals

Externally studentized residuals with 95% simultaneous limits (Bonferroni) in red, individual limits in green.

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	10.152866	3.049517	3.33	0.0009*
pH CLP	-0.158684	0.155358	-1.02	0.3076
FIC102	0.0007034	0.000227	3.09	0.0021*
FIC105	-0.000932	0.00072	-1.29	0.1961
FIC107	-0.000897	0.000629	-1.42	0.1551
FIC112	0.0013635	0.00071	1.92	0.0556
FIC401	-0.000131	0.000032	-4.09	<.0001*
FIC175F_L	-0.000229	0.000341	-0.67	0.5028
FIC8_401	-0.001234	0.000565	-2.18	0.0296*
FIC108	-3.661e-5	0.000508	-0.07	0.9426
FIC155	-0.000028	0.000506	-0.06	0.9560
PHIC401	0.0640425	0.148641	0.43	0.6668
FIC104	-7.544e-5	0.001039	-0.07	0.9421
FIC156	0.0001568	0.00127	0.12	0.9018
FIC158	-1.788e-5	0.000506	-0.04	0.9718
FIC181	-0.001214	0.00253	-0.48	0.6317
FIC310	-0.025422	0.05353	-0.47	0.6351
FIC2_131	0.0013826	0.00258	0.54	0.5923

Figure 2.11: Least Squares all variables



Externally studentized residuals with 95% simultaneous limits (Bonferroni) in red, individual limits in green.

Effect Summary

Source	Logworth	PValue
FIC401	4.591	0.00003
FIC112	3.510	0.00031
FIC102	3.186	0.00065
FIC107	2.652	0.00223
FIC8_401	2.225	0.00595
FIC105	1.788	0.01631
FIC175F_L	1.344	0.04527
pH CLP	1.207	0.06208

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
FIC102	1	1	0.08999105	11.7884	0.0007*
FIC105	1	1	0.04437978	5.8136	0.0163*
FIC107	1	1	0.07221659	9.4600	0.0022*
FIC112	1	1	0.10094843	13.2238	0.0003*
FIC401	1	1	0.13813405	18.0949	<.0001*
FIC8_401	1	1	0.05830334	7.6375	0.0060*
pH CLP	1	1	0.02670761	3.4986	0.0621
FIC175F_L	1	1	0.03077529	4.0314	0.0453*

Summary of Fit

RSquare	0.989652	AICc	-911.483
RSquare Adj	0.989465	BIC	-870.822
Root Mean Square Error	0.087372		
Mean of Response	4.36245		
Observations (or Sum Wgts)	453		

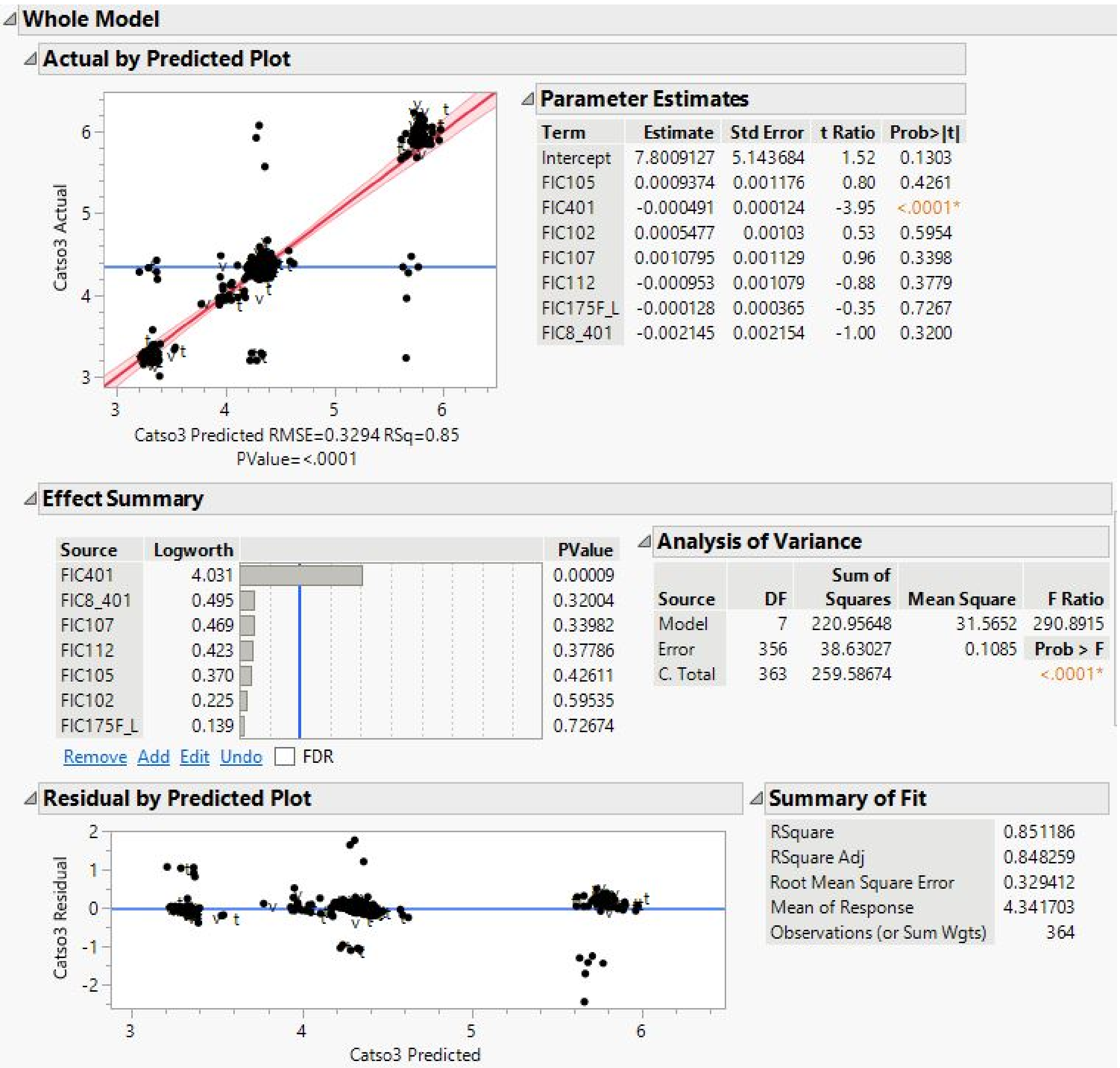
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	8	324.14575	40.5182	5307.704
Error	444	3.38943	0.0076	**Prob > F**
C. Total	452	327.53518		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	9.9063789	1.851375	5.35	<.0001*
FIC102	0.00074	0.000216	3.43	0.0007*
FIC105	-0.000946	0.000392	-2.41	0.0163*
FIC107	-0.000885	0.000288	-3.08	0.0022*
FIC112	0.00128	0.000352	3.64	0.0003*
FIC401	-0.000132	0.000031	-4.25	<.0001*
FIC8_401	-0.001336	0.000483	-2.76	0.0060*
pH CLP	-0.123504	0.066029	-1.87	0.0621
FIC175F_L	-0.000253	0.000126	-2.01	0.0453*

Figure 2.12: Least Squares reduced variables



Whole Model

Actual by Predicted Plot

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	7.8009127	5.143684	1.52	0.1303
FIC105	0.0009374	0.001176	0.80	0.4261
FIC401	-0.000491	0.000124	-3.95	<.0001*
FIC102	0.0005477	0.00103	0.53	0.5954
FIC107	0.0010795	0.001129	0.96	0.3398
FIC112	-0.000953	0.001079	-0.88	0.3779
FIC175F_L	-0.000128	0.000365	-0.35	0.7267
FIC8_401	-0.002145	0.002154	-1.00	0.3200

Effect Summary

Source	Logworth		PValue
FIC401	4.031		0.00009
FIC8_401	0.495		0.32004
FIC107	0.469		0.33982
FIC112	0.423		0.37786
FIC105	0.370		0.42611
FIC102	0.225		0.59535
FIC175F_L	0.139		0.72674

Remove Add Edit Undo □ FDR

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	7	220.95648	31.5652	290.8915
Error	356	38.63027	0.1085	Prob > F
C. Total	363	259.58674		<.0001*

Residual by Predicted Plot

Summary of Fit

RSquare	0.851186
RSquare Adj	0.848259
Root Mean Square Error	0.329412
Mean of Response	4.341703
Observations (or Sum Wgts)	364

Figure 2.13: Least Squares reduced variables without ph

3
Models

The main focus of the project is to create a predictive modelling for CatSO3. One step ahead was the remove certain features from the prediction. For example brand, reblend and other features that has correlation and noise were removed from our list of features. Now that we have left with core data set with 8 features and a response variable. These features are FIC102, FIC105, FIC107, FIC112, FIC401, FIC175F-L, FIC8-401 and ph. Now we choose our models to fit and later on continue to compare them. In the upcoming subsections, in each section we will present a model explanation and later on our application on that model.

Predictive modeling is the process of taking known results and developing a model that can predict values for new occurrences. It uses historical data to predict future events. There are many different types of predictive modeling techniques including ANOVA, linear regression (ordinary least squares), logistic regression, ridge regression, time series, decision trees, neural networks, and many more. Selecting the correct predictive modeling technique at the start of your project can save a lot of time. Choosing the incorrect modeling technique can result in inaccurate predictions and residual plots that experience non-constant variance and/or mean. [6] In our case since both response and features are continuous variables, we will look into regression techniques, decision trees, machine learning and deep learning techniques. In this project we will specifically focus on 4 models, which are: Generalized Linear Model (GLR), Neural Network model (NN), Bootstrap Forest model (BF), Support Vector Machines (SVM).

For Generalized Linear Model we will use two different variable selection techniques, including shrinkage techniques, that specifically address modeling correlated and high-dimensional data. Two of these techniques, the Lasso and the Elastic Net, perform variable selection as part of the modeling procedure. Even for small data sets with little or no correlation, including designed experiments, the Lasso and Elastic Net are useful. They can be used to build predictive models or to select variables for model reduction or for future study. Modeling techniques such as the Elastic Net and the Lasso are particularly useful for large data sets, where collinearity is typically a problem. In addition, modern data sets often include more variables than observations. This situation is sometimes referred to as the $p > n$ problem, where n is the number of observations and p is the number of predictors. Such

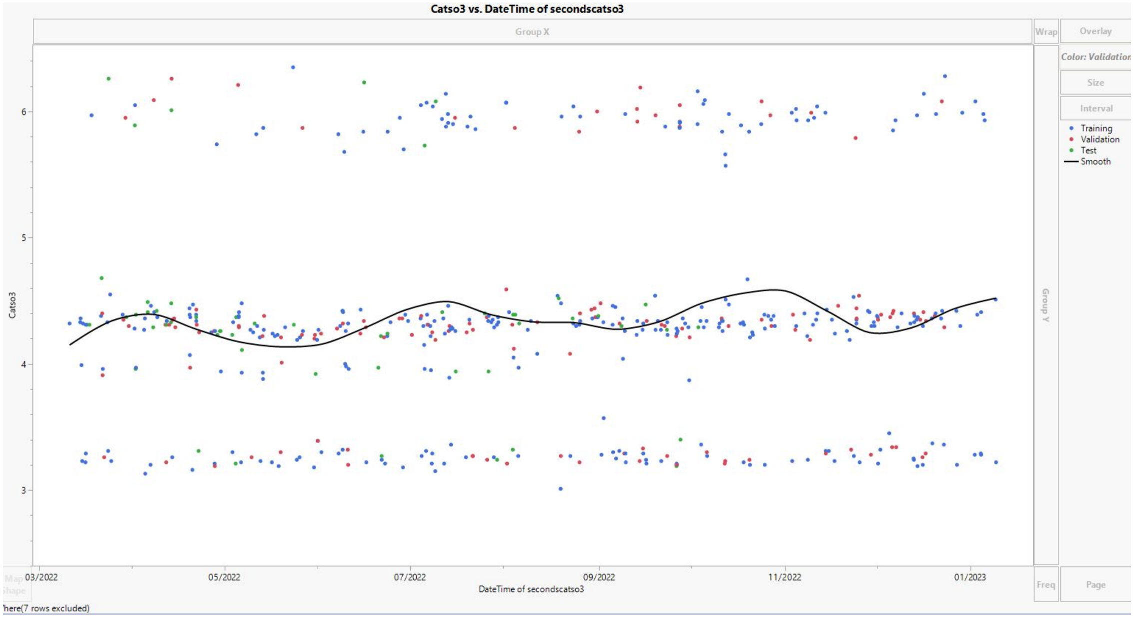


Figure 3.1: Randomly chosen Validation data

data sets require variable selection if traditional modeling techniques are to be used. The Elastic Net and Lasso are relatively recent techniques [7]. Both techniques does penalize the size of the model coefficients, and continuous to a shrinkage in the end. The shrinkage is determined by an adjustable variable. An optimal shrinkage is decided by one of the several validation methods but in our case we have separated data into Training, Validation and Tests sets while using predictive modelling. Also we took special precautions while creating Validation data set. Randomization of validation data set is a good option but after completing GLR model we decided to create a special case of Validation data set to see if the models improve. In the model section section we tested two different validation data sets to see the changes. First validation data set is randomly selected between all the rows Figure 3.1. The second validation data set has been created and grouped by brand and also as we take one brand as our sample, first 15 percent become our validation group following 70 percent become training and last 15 percent become our test group per brand Figure 3.2.

3.1 Generalized Linear Modelling - Lasso and Elasticnet

A generalized linear model (GLM) generalizes linear regression by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value. We believe that error distribution of responses behave normal distribution, so it is a good choice to select GLM. How ever we also would like to show variable reduction via Lasso penalization. In the previous part we presented that variables are reduced to 8 predictor. By using GLM-Lasso we will also

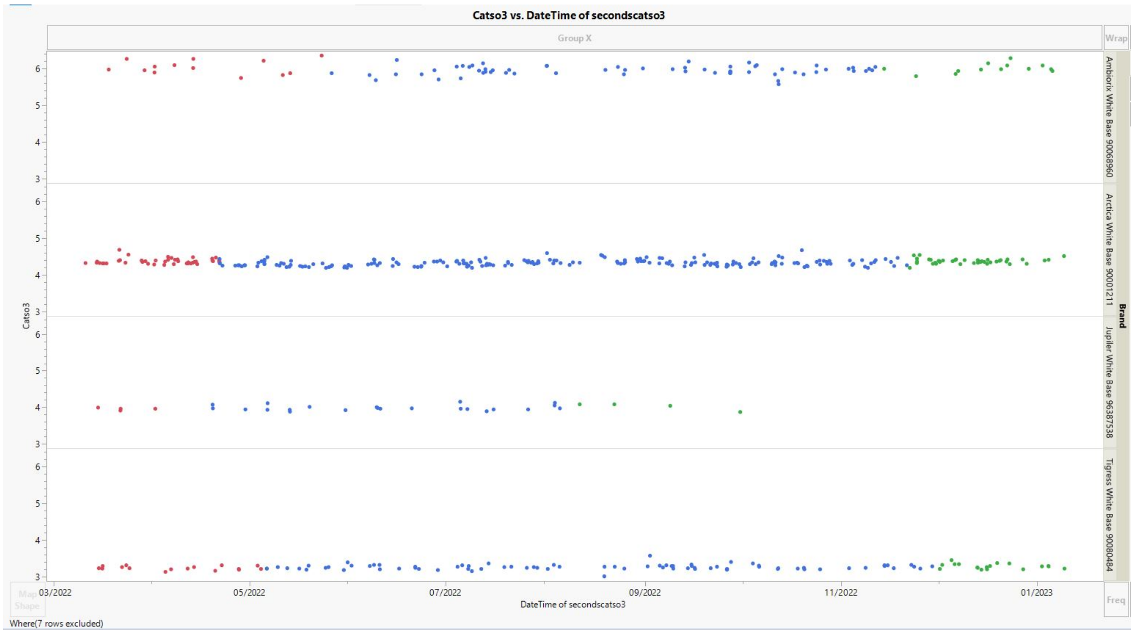


Figure 3.2: Validation data created by brand categorization

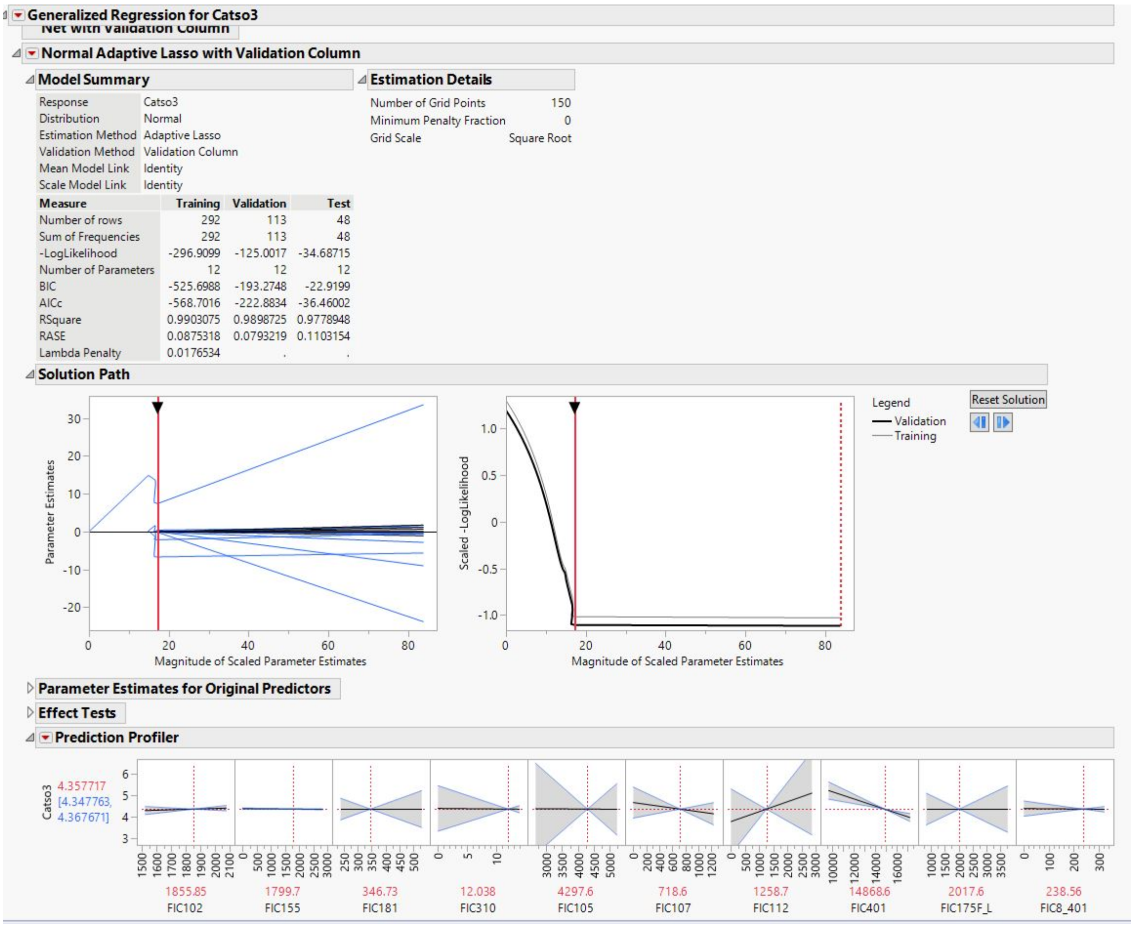
prove that we can indeed reduce the dimentionality of the variables into exact same number with same features. There are few aspects of Lasso we should consider while we do the modelling. When several variables are highly correlated, Lasso tends to select only one variable from that group. When the number of variables, p, exceeds the number of observations, n, the Lasso selects at most n predictors. This is in fact our case that features have highly correlated.

Regularization is a method for solving ill-posed problems or problems of models overfitting data. The method involves introducing additional information to a model in the form of a penalty. In terms of Elasticnet and LASSO, the penalty imposes a shrinkage on the coefficient estimates of ordinary least squares. This penalty controls the instability found in the least squares model with nonorthogonal matrices. Generally, for the L_p regularization term we have $L_p = (\sum_i ||\beta_i||^p)^{\frac{1}{p}}$. Elasticnet and LASSO deal with the L_2 and L_1 penalties respectively. Regularization is used in preference over other common methods of determining the best linear model, such as best subset selection and stepwise subset selection. Elastic net is the same as lasso when α = 1. For other values of α, the penalty term L_p interpolates between the L1 norm of β and the squared L2 norm of β. As α shrinks toward 0, elastic net approaches ridge regression.[8]

$$L_p = \sum_{j=1}^{p} \left(\frac{(1-\alpha)}{2} \beta_j^2 \ + \alpha |\beta_j| \right). \ [8]$$

For a nonnegative value of λ, algorithm solves the problem:

$$min_(\beta_0, \beta)(\tfrac{1}{N} SSE(\beta_0, \beta) + \lambda L_p)[8]$$



Generalized Regression for Catso3

Net with Validation Column

Normal Adaptive Lasso with Validation Column

Model Summary

Response	Catso3
Distribution	Normal
Estimation Method	Adaptive Lasso
Validation Method	Validation Column
Mean Model Link	Identity
Scale Model Link	Identity

Measure	Training	Validation	Test
Number of rows	292	113	48
Sum of Frequencies	292	113	48
-LogLikelihood	-296.9099	-125.0017	-34.68715
Number of Parameters	12	12	12
BIC	-525.6988	-193.2748	-22.9199
AICc	-568.7016	-222.8834	-36.46002
RSquare	0.9903075	0.9898725	0.9778948
RASE	0.0875318	0.0793219	0.1103154
Lambda Penalty	0.0176534	.	.

Estimation Details

Number of Grid Points	150
Minimum Penalty Fraction	0
Grid Scale	Square Root

Solution Path

Parameter Estimates for Original Predictors

Effect Tests

Prediction Profiler

Figure 3.3: GLR Lasso all variables

Minimizing the λ-penalized deviance is equivalent to maximizing the λ-penalized loglikelihood. N is the number of observations.λ is a nonnegative regularization parameter corresponding to one value of Lambda.The parameters β0 and β are a scalar and a vector of length p, respectively. [9]

After looking at the Figure 3.3 and Figure 3.4 we can see that for GLR Lasso, it reduces the all variable into the version of selected variables as per-se.

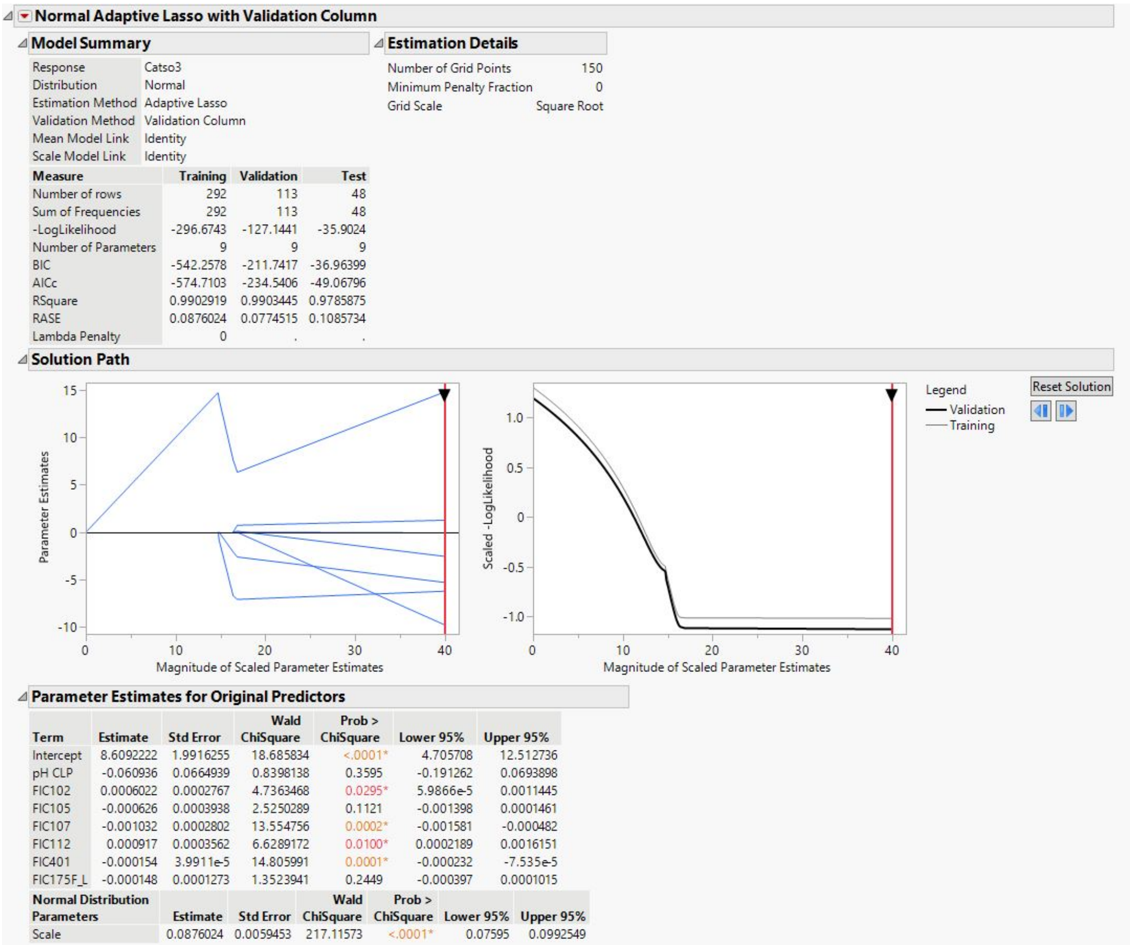
Let's analyse the results of Normal adaptive lasso's two options we have, one with the all features and the other is selected features. As we compare selected versus all variables Figure 3.3 and Figure 3.4, we will see that all feature solution has lower R square value for validation and test sets. It reduced the feature size to 12 including 3 more extra features including FIC155,FIC310 but it still looks like overfitting comparing to the selected features solution. AIC and BIC scores are low but comparing to the selected variable not so low. The little difference between R squares, AIC, BIC and RASE values suggests that we should go with the selected feature model, even though statistically both models are valid.

By following above argument, you will see the same results of all features versus selected features results of generalized Linear Regression with adaptive elasticnet penalization between Figure 3.8 and Figure 3.10. Same argument and same weaknesses for both penalizations method. That will suggests us that feature selection for creating a model, even with penalization, sometimes makes difference, even it is a small difference. That is why process knowledge helps us to create a better model by removing unnecessary noises from the model that we want to create.

There is also the comparison between random validation set versus selected variables including adjusted validation data set. To make it easy to compare we should compare the same used models, meaning for example if it is generalized linear regression with lasso penalization, we should take two model made by same selected features but compare the parts where the validation data set is different. First lets look into Figure 3.4 and Figure 3.6. Only difference between these models are the validation data sets. In this case their R square looks very close to each other but we can also compare them via AIC and BIC squares. In our case the lower the score is the better for our analysis. In this case the model which uses adapted validation set by brand has lower AIC and BIC value. By that means we would like to take into the consideration, the one model that uses adapted (by brand) validation data set.For the models that uses Elastic net penalization method Figure 3.10 and Figure 3.12 you can see from the AIC and BIC values same result applies. That is why validation data set created by brand adaptation is important in our analysis.

For the models that uses Elastic net penalization method Figure 3.10 and Figure 3.12 you can see from the AIC and BIC values, same result applies. In our analysis we would like to take the ones with lower AIC and BIC valued models, even though F test suggest that both model is valid. That is why validation data set created by brand adaptation is important in our analysis.

Both models of Generalized Linear Regression with Lasso and Adaptive Elastic net penalization with selected feature gives us good results. now we are considering both model which uses adapted validation sets. R square values are close to %99 with a small differences between test set R square values. We can also compare them visually by looking at the predicted versus real responses graphs in Figure 3.7 and Figure 3.13. Further more JMP allows us to use prediction profilers interactively. These profilers help us to see how much change will occur in the other predictor if the values of one predictor changes. This feature also useful to compare the slopes of the features as well. The more slope one predictor have meaning more effect on the response variable. This could be extrapolated via "Bvalues magnitude as well but visually seeing it and interacting with it gives us the decision choice of the



Normal Adaptive Lasso with Validation Column

Model Summary

Response	Catso3
Distribution	Normal
Estimation Method	Adaptive Lasso
Validation Method	Validation Column
Mean Model Link	Identity
Scale Model Link	Identity

Measure	Training	Validation	Test
Number of rows	292	113	48
Sum of Frequencies	292	113	48
-LogLikelihood	-296.6743	-127.1441	-35.9024
Number of Parameters	9	9	9
BIC	-542.2578	-211.7417	-36.96399
AICc	-574.7103	-234.5406	-49.06796
RSquare	0.9902919	0.9903445	0.9785875
RASE	0.0876024	0.0774515	0.1085734
Lambda Penalty	0	.	.

Estimation Details

Number of Grid Points	150
Minimum Penalty Fraction	0
Grid Scale	Square Root

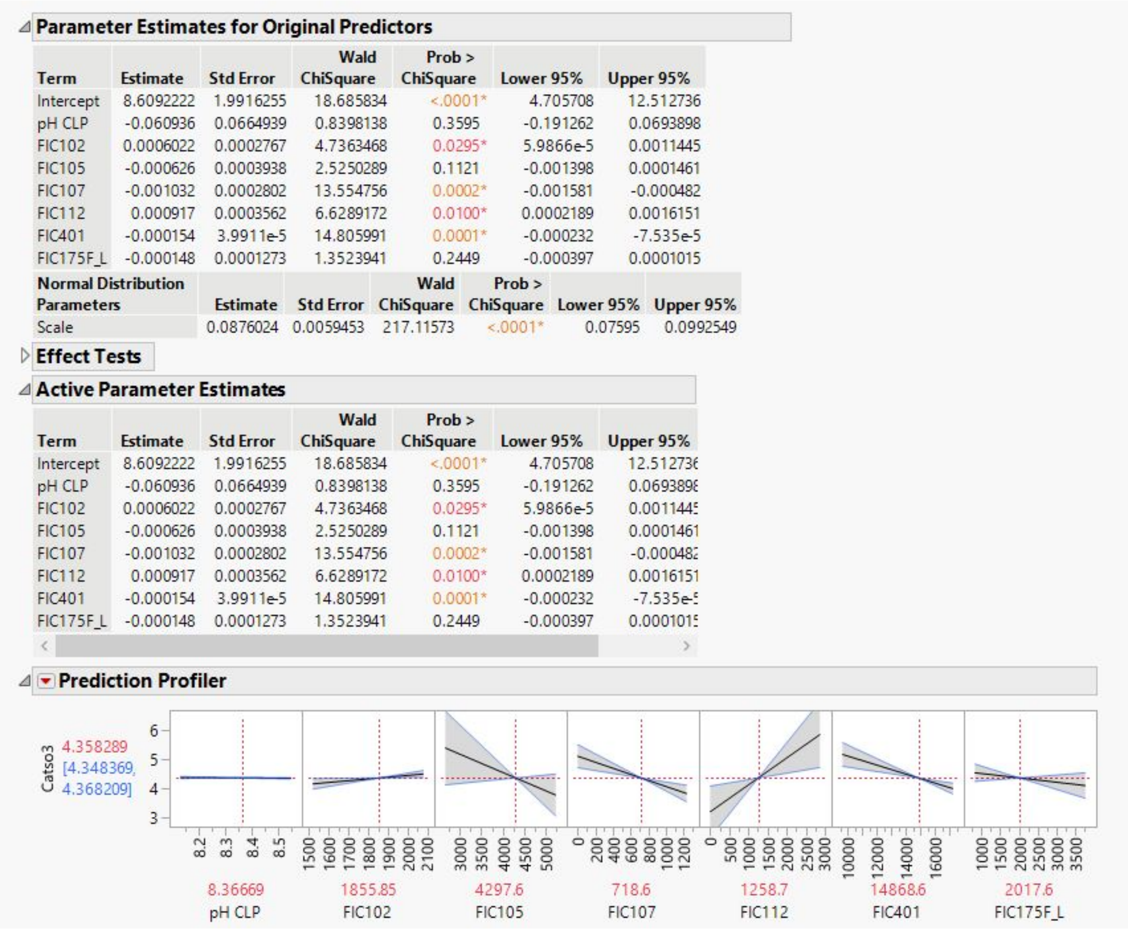
Solution Path

Parameter Estimates for Original Predictors

Term	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
Intercept	8.6092222	1.9916255	18.685834	<.0001*	4.705708	12.512736
pH CLP	-0.060936	0.0664939	0.8398138	0.3595	-0.191262	0.0693898
FIC102	0.0006022	0.0002767	4.7363468	0.0295*	5.9866e-5	0.0011445
FIC105	-0.000626	0.0003938	2.5250289	0.1121	-0.001398	0.0001461
FIC107	-0.001032	0.0002802	13.554756	0.0002*	-0.001581	-0.000482
FIC112	0.000917	0.0003562	6.6289172	0.0100*	0.0002189	0.0016151
FIC401	-0.000154	3.9911e-5	14.805991	0.0001*	-0.000232	-7.535e-5
FIC175F_L	-0.000148	0.0001273	1.3523941	0.2449	-0.000397	0.0001015

Normal Distribution Parameters	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
Scale	0.0876024	0.0059453	217.11573	<.0001*	0.07595	0.0992549

Figure 3.4: GLR Lasso selected variables 1



Parameter Estimates for Original Predictors

Term	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
Intercept	8.6092222	1.9916255	18.685834	<.0001*	4.705708	12.512736
pH CLP	-0.060936	0.0664939	0.8398138	0.3595	-0.191262	0.0693898
FIC102	0.0006022	0.0002767	4.7363468	0.0295*	5.9866e-5	0.0011445
FIC105	-0.000626	0.0003938	2.5250289	0.1121	-0.001398	0.0001461
FIC107	-0.001032	0.0002802	13.554756	0.0002*	-0.001581	-0.000482
FIC112	0.000917	0.0003562	6.6289172	0.0100*	0.0002189	0.0016151
FIC401	-0.000154	3.9911e-5	14.805991	0.0001*	-0.000232	-7.535e-5
FIC175F_L	-0.000148	0.0001273	1.3523941	0.2449	-0.000397	0.0001015

Normal Distribution Parameters	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
Scale	0.0876024	0.0059453	217.11573	<.0001*	0.07595	0.0992549

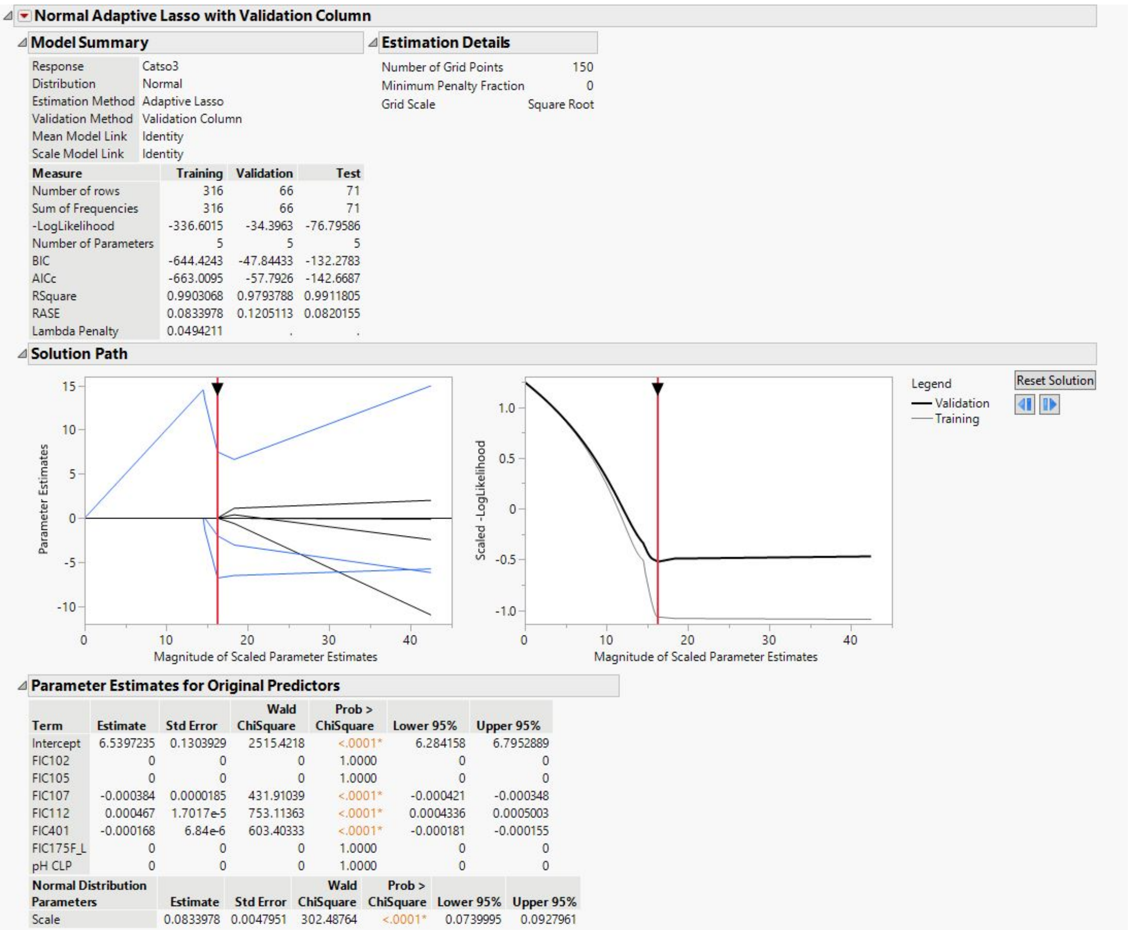
Effect Tests

Active Parameter Estimates

Term	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
Intercept	8.6092222	1.9916255	18.685834	<.0001*	4.705708	12.51273
pH CLP	-0.060936	0.0664939	0.8398138	0.3595	-0.191262	0.069389
FIC102	0.0006022	0.0002767	4.7363468	0.0295*	5.9866e-5	0.001144
FIC105	-0.000626	0.0003938	2.5250289	0.1121	-0.001398	0.0001461
FIC107	-0.001032	0.0002802	13.554756	0.0002*	-0.001581	-0.00048
FIC112	0.000917	0.0003562	6.6289172	0.0100*	0.0002189	0.0016151
FIC401	-0.000154	3.9911e-5	14.805991	0.0001*	-0.000232	-7.535e-
FIC175F_L	-0.000148	0.0001273	1.3523941	0.2449	-0.000397	0.000101

Prediction Profiler

Figure 3.5: GLR Lasso selected variables 2



Normal Adaptive Lasso with Validation Column

Model Summary

Response	Catso3
Distribution	Normal
Estimation Method	Adaptive Lasso
Validation Method	Validation Column
Mean Model Link	Identity
Scale Model Link	Identity

Measure	Training	Validation	Test
Number of rows	316	66	71
Sum of Frequencies	316	66	71
-LogLikelihood	-336.6015	-34.3963	-76.79586
Number of Parameters	5	5	5
BIC	-644.4243	-47.84433	-132.2783
AICc	-663.0095	-57.7926	-142.6687
RSquare	0.9903068	0.9793788	0.9911805
RASE	0.0833978	0.1205113	0.0820155
Lambda Penalty	0.0494211	.	.

Estimation Details

Number of Grid Points	150
Minimum Penalty Fraction	0
Grid Scale	Square Root

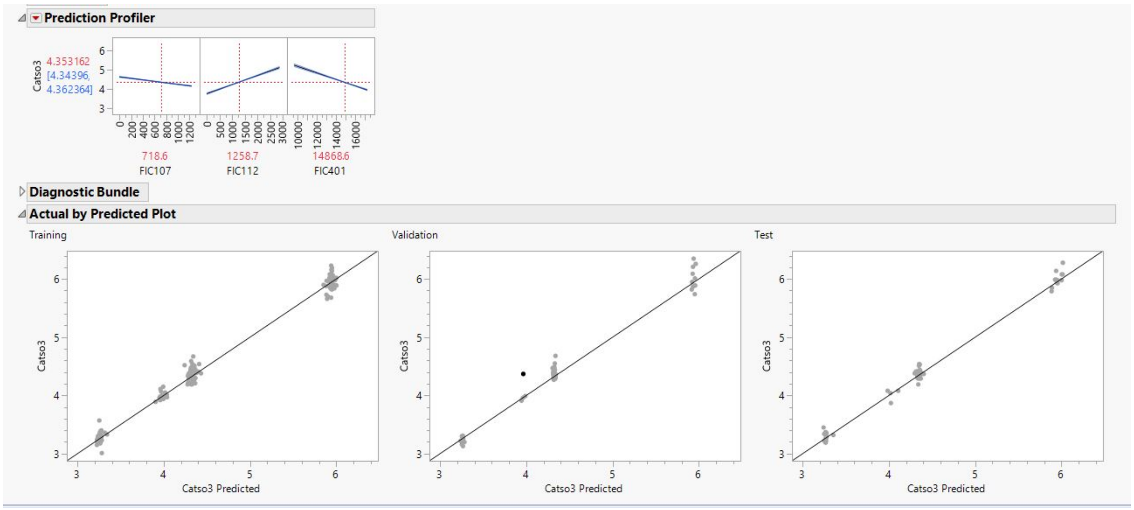
Solution Path

Parameter Estimates for Original Predictors

Term	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
Intercept	6.5397235	0.1303929	2515.4218	<.0001*	6.284158	6.7952889
FIC102	0	0	0	1.0000	0	0
FIC105	0	0	0	1.0000	0	0
FIC107	-0.000384	0.0000185	431.91039	<.0001*	-0.000421	-0.000348
FIC112	0.000467	1.7017e-5	753.11363	<.0001*	0.0004336	0.0005003
FIC401	-0.000168	6.84e-6	603.40333	<.0001*	-0.000181	-0.000155
FIC175F_L	0	0	0	1.0000	0	0
pH CLP	0	0	0	1.0000	0	0

Normal Distribution Parameters	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
Scale	0.0833978	0.0047951	302.48764	<.0001*	0.0739995	0.0927961

Figure 3.6: GLR Lasso selected variables with grouped brands validation data set 1

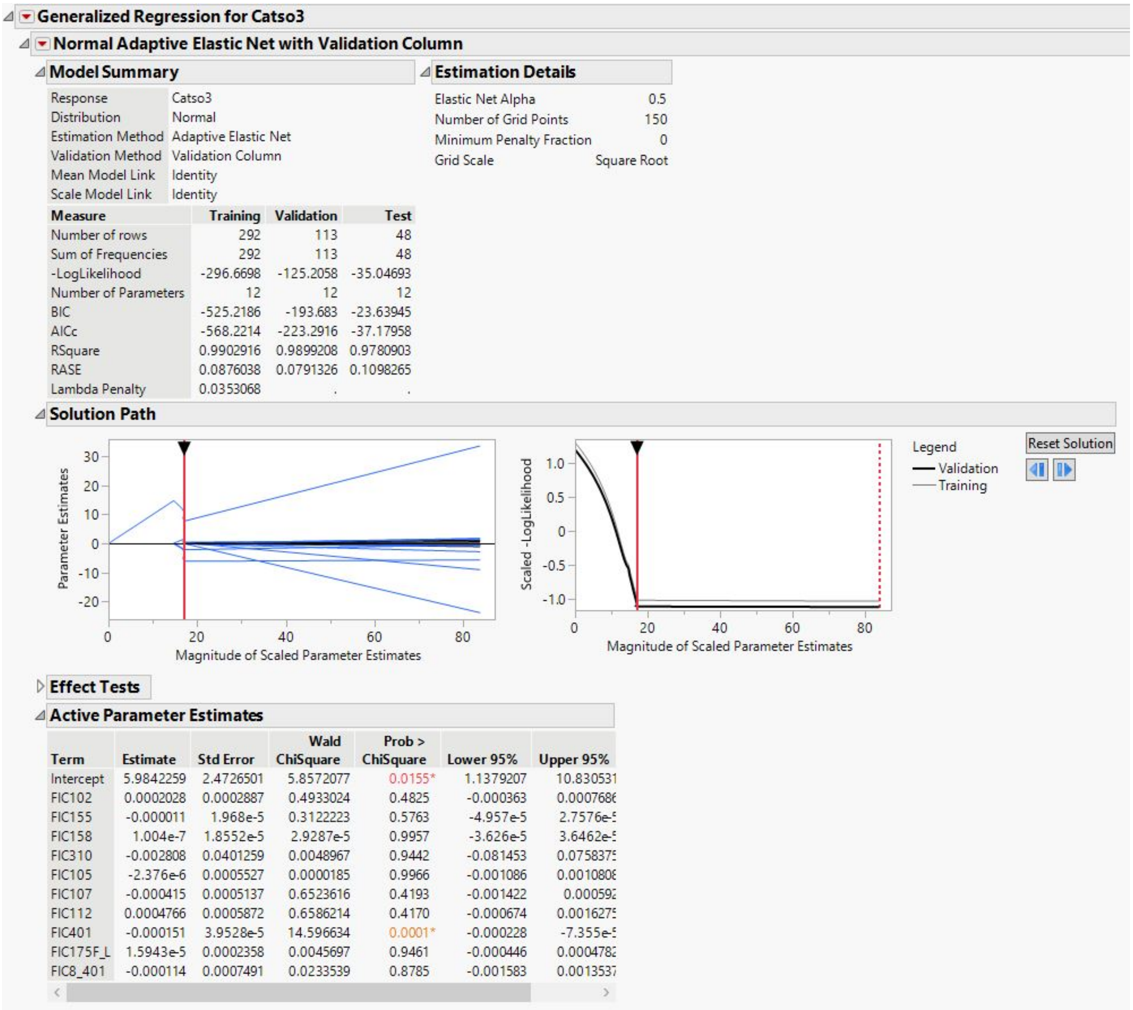


Prediction Profiler

Diagnostic Bundle

Actual by Predicted Plot

Figure 3.7: GLR Lasso selected variables with grouped brands validation data set 2



Generalized Regression for Catso3

Normal Adaptive Elastic Net with Validation Column

Model Summary

Response	Catso3
Distribution	Normal
Estimation Method	Adaptive Elastic Net
Validation Method	Validation Column
Mean Model Link	Identity
Scale Model Link	Identity

Measure	Training	Validation	Test
Number of rows	292	113	48
Sum of Frequencies	292	113	48
-LogLikelihood	-296.6698	-125.2058	-35.04693
Number of Parameters	12	12	12
BIC	-525.2186	-193.683	-23.63945
AICc	-568.2214	-223.2916	-37.17958
RSquare	0.9902916	0.9899208	0.9780903
RASE	0.0876038	0.0791326	0.1098265
Lambda Penalty	0.0353068	.	.

Estimation Details

Elastic Net Alpha	0.5
Number of Grid Points	150
Minimum Penalty Fraction	0
Grid Scale	Square Root

Solution Path

Effect Tests

Active Parameter Estimates

Term	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
Intercept	5.9842259	2.4726501	5.8572077	0.0155*	1.1379207	10.830531
FIC102	0.0002028	0.0002887	0.4933024	0.4825	-0.000363	0.0007686
FIC155	-0.000011	1.968e-5	0.3122223	0.5763	-4.957e-5	2.7576e-5
FIC158	1.004e-7	1.8552e-5	2.9287e-5	0.9957	-3.626e-5	3.6462e-5
FIC310	-0.002808	0.0401259	0.0048967	0.9442	-0.081453	0.0758375
FIC105	-2.376e-6	0.0005527	0.0000185	0.9966	-0.001086	0.0010808
FIC107	-0.000415	0.0005137	0.6523616	0.4193	-0.001422	0.000592
FIC112	0.0004766	0.0005872	0.6586214	0.4170	-0.000674	0.0016275
FIC401	-0.000151	3.9528e-5	14.596634	0.0001*	-0.000228	-7.355e-5
FIC175F_L	1.5943e-5	0.0002358	0.0045697	0.9461	-0.000446	0.0004782
FIC8_401	-0.000114	0.0007491	0.0233539	0.8785	-0.001583	0.0013537

Figure 3.8: GLR Elastic all variables 1

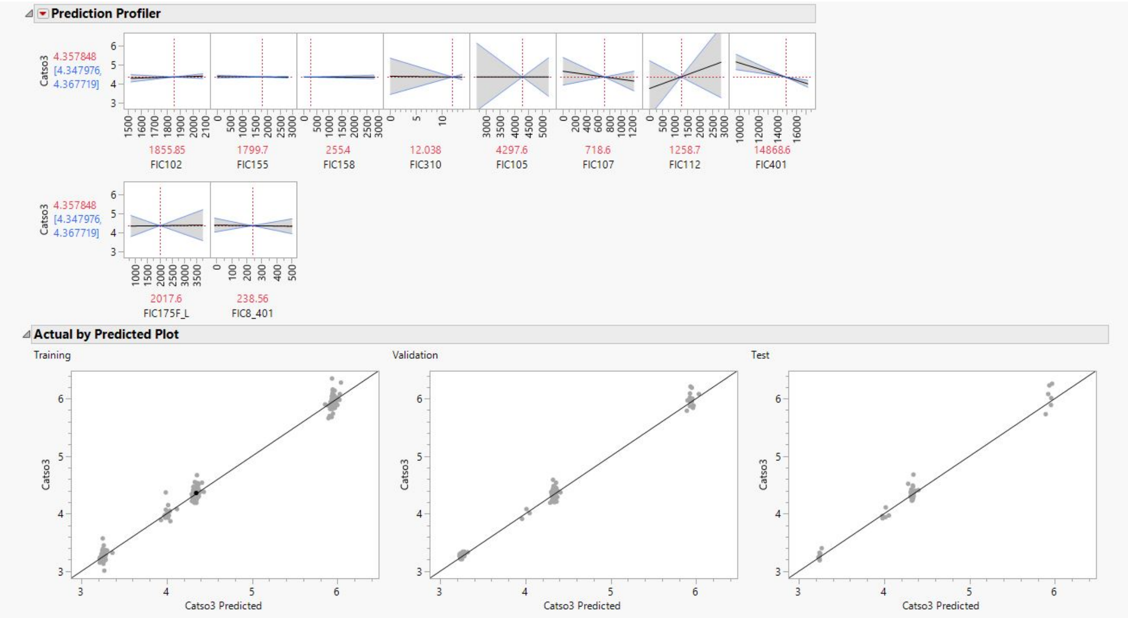
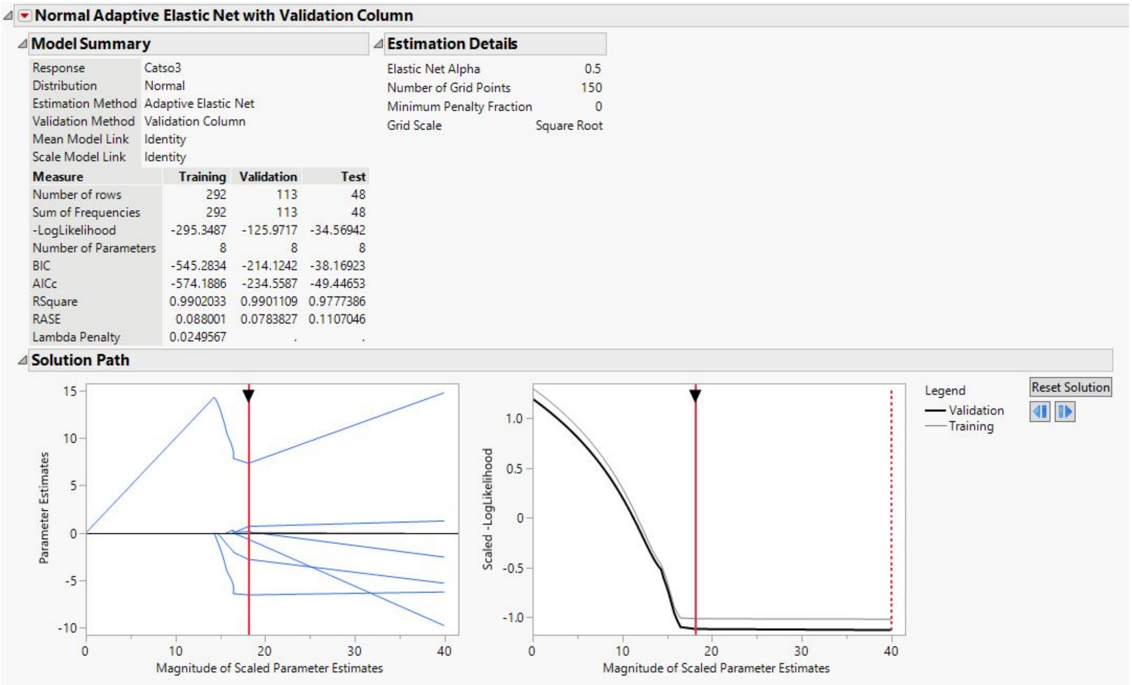


Figure 3.9: GLR Elastic all variables 2



Normal Adaptive Elastic Net with Validation Column

Model Summary

Response	Catso3
Distribution	Normal
Estimation Method	Adaptive Elastic Net
Validation Method	Validation Column
Mean Model Link	Identity
Scale Model Link	Identity

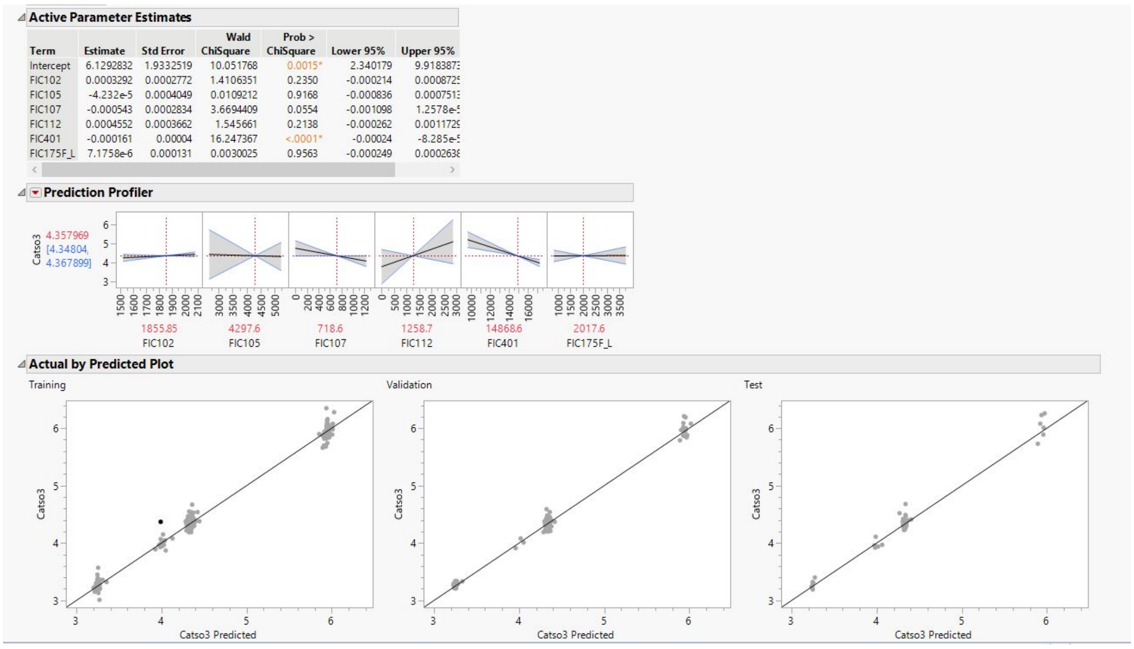
Measure	Training	Validation	Test
Number of rows	292	113	48
Sum of Frequencies	292	113	48
-LogLikelihood	-295.3487	-125.9717	-34.56942
Number of Parameters	8	8	8
BIC	-545.2834	-214.1242	-38.16923
AICc	-574.1886	-234.5587	-49.44653
RSquare	0.9902033	0.9901109	0.9777386
RASE	0.088001	0.0783827	0.1107046
Lambda Penalty	0.0249567	.	.

Estimation Details

Elastic Net Alpha	0.5
Number of Grid Points	150
Minimum Penalty Fraction	0
Grid Scale	Square Root

Solution Path

Figure 3.10: GLR Elastic selected variables 1



Active Parameter Estimates

Term	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
Intercept	6.1292832	1.9332519	10.051768	0.0015*	2.340179	9.918387
FIC102	0.0003292	0.0002772	1.4106351	0.2350	-0.000214	0.000872
FIC105	-4.232e-5	0.0004049	0.0109212	0.9168	-0.000836	0.000751
FIC107	-0.000543	0.0002834	3.6694409	0.0554	-0.001098	1.2578e-
FIC112	0.0004552	0.0003662	1.545661	0.2138	-0.000262	0.001172
FIC401	-0.000161	0.00004	16.247367	<.0001*	-0.00024	-8.285e-
FIC175F_L	7.1758e-6	0.000131	0.0030025	0.9563	-0.000249	0.000263

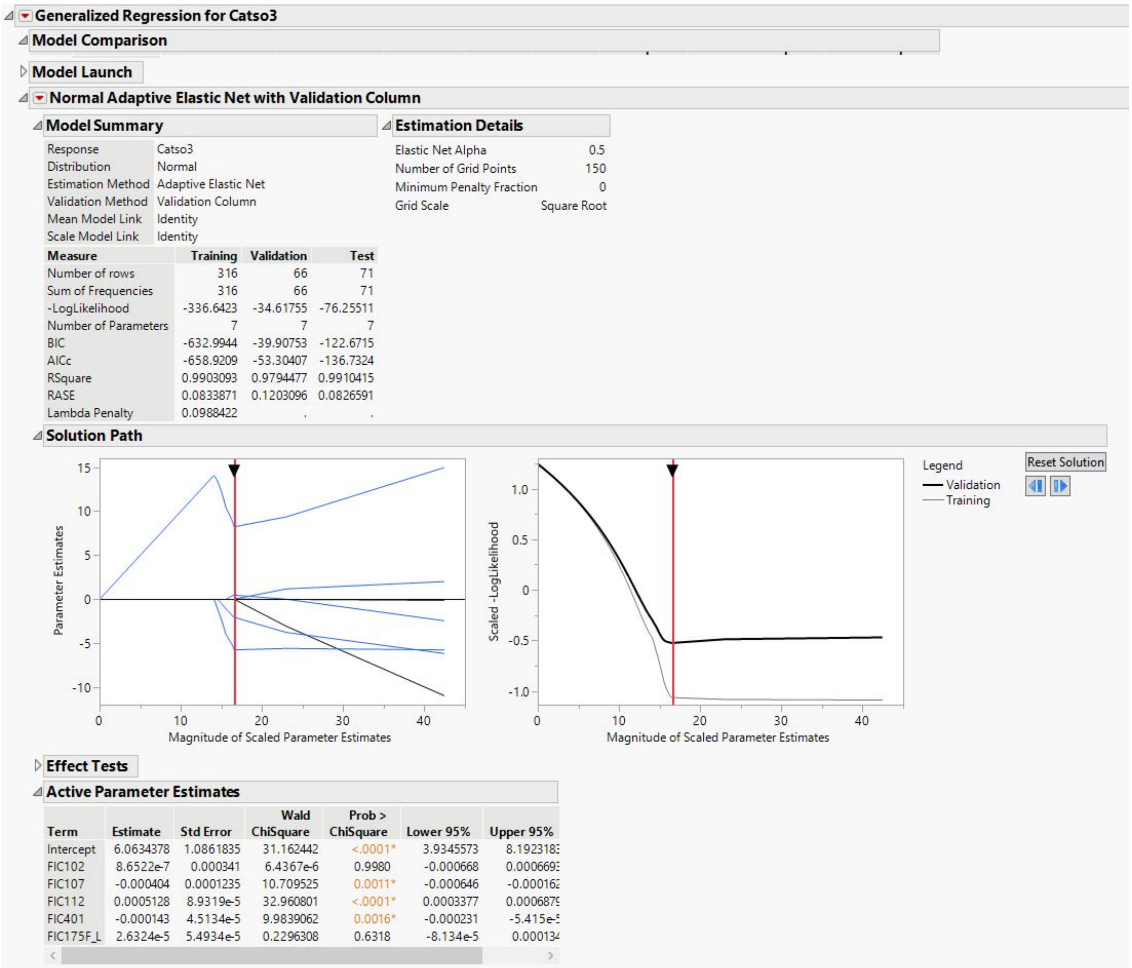
Figure 3.11: GLR Elastic selected variables 2

production parameters which we should keep an eye on. These profilers will help production to create confidence intervals for how much they can deviate on one feature movement. This is a really useful tool for process engineers and will be put to use once the model is active on the line. According to flow rates of features it might be possible to exchange ingredients for economic reasons. Although there are other quality checks that is out of this project scope, it might be used by research and development departments to come up with new ideas.

3.2 Support Vector Machines

A support vector machine (SVM) model is a supervised learning algorithm that is used to predict or classify new observations. A model is fit on a set of training data where the responses are known. Then, the model is used to predict the responses of new observations. When the response is continuous, the models that are fit are known as support vector regression (SVR) models. In a typical regression problem, the goal is to fit a model that minimizes the error between a predicted response and the actual response. In an SVR problem, the goal is to fit a model such that the error between a predicted response and the actual response falls within a range of -ε to ε. This provides a more flexible fit. In our model, ε is equal to 0.1. The SVR algorithm doubles the data by creating two classes, Y + ε and Y - ε. Then the same algorithm that is used for the classification problem is also used for the prediction (SVR) problem. A linear kernel is a simple dot product between two input vectors.

SVM (ε-SVM) regression, which is also known as L1 loss. In ε-SVM regression, the set of training data includes predictor variables and observed response values. The goal is to find a function f(x) that deviates from y_n by value ε for each training point x. Data where x_n is a multivariate set of N observations with observed response values y_n.



Generalized Regression for Catso3

Model Comparison

Model Launch

Normal Adaptive Elastic Net with Validation Column

Model Summary

Response	Catso3
Distribution	Normal
Estimation Method	Adaptive Elastic Net
Validation Method	Validation Column
Mean Model Link	Identity
Scale Model Link	Identity

Measure	Training	Validation	Test
Number of rows	316	66	71
Sum of Frequencies	316	66	71
-LogLikelihood	-336.6423	-34.61755	-76.25511
Number of Parameters	7	7	7
BIC	-632.9944	-39.90753	-122.6715
AICc	-658.9209	-53.30407	-136.7324
RSquare	0.9903093	0.9794477	0.9910415
RASE	0.0833871	0.1203096	0.0826591
Lambda Penalty	0.0988422	.	.

Estimation Details

Elastic Net Alpha	0.5
Number of Grid Points	150
Minimum Penalty Fraction	0
Grid Scale	Square Root

Solution Path

Effect Tests

Active Parameter Estimates

Term	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
Intercept	6.0634378	1.0861835	31.162442	<.0001*	3.9345573	8.192318
FIC102	8.6522e-7	0.000341	6.4367e-6	0.9980	-0.000668	0.000669
FIC107	-0.000404	0.0001235	10.709525	0.0011*	-0.000646	-0.00016
FIC112	0.0005128	8.9319e-5	32.960801	<.0001*	0.0003377	0.000687
FIC401	-0.000143	4.5134e-5	9.9839062	0.0016*	-0.000231	-5.415e-
FIC175F_L	2.6324e-5	5.4934e-5	0.2296308	0.6318	-8.134e-5	0.00013

Figure 3.12: GLR Elastic with selected variables including adjusted validation data set 1

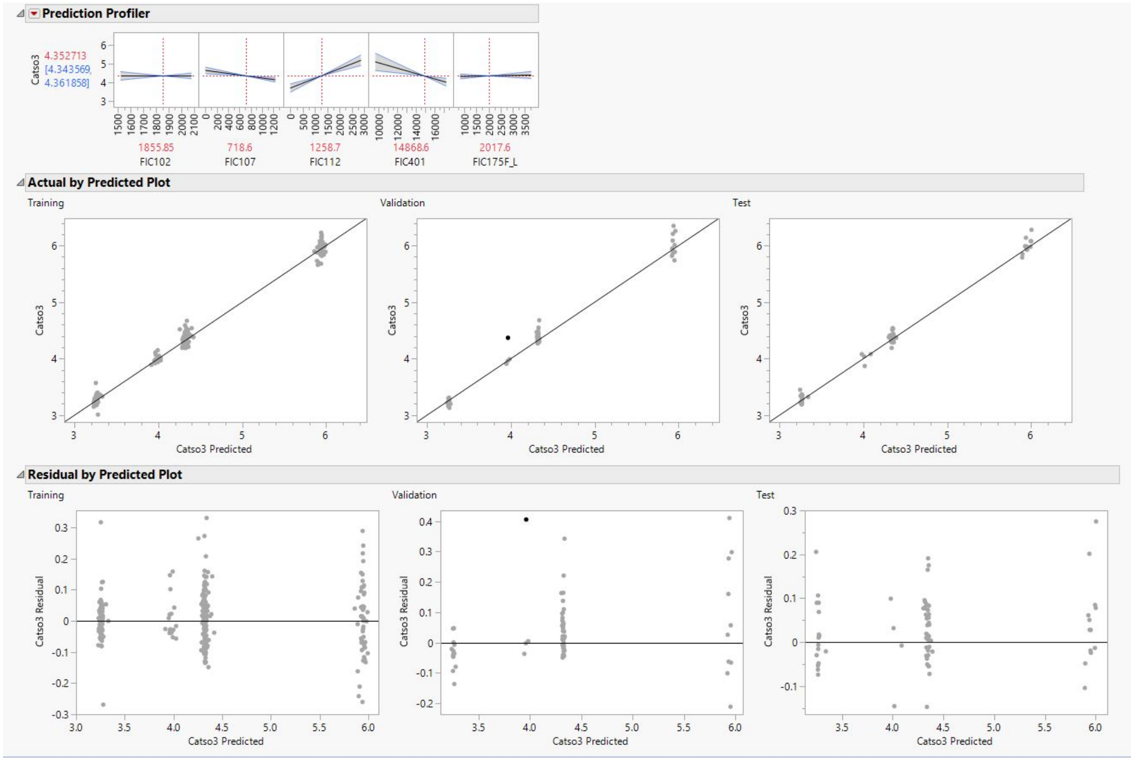


Figure 3.13: GLR Elastic with selected variables including adjusted validation data set 2

$$f(x) = x'\beta + b$$

and ensure that it is as flat as possible, find f(x) with the minimal norm value (β′β). This is formulated as a convex optimization problem to minimize.

$$J(\beta) = \frac{1}{2}\beta'\beta$$

subject to all residuals having a value less than ε ; The β parameter can be completely described as a linear combination of the training observations using the equation.

$$\beta = \sum_{n=1}^{N}(a_n - a_n^*)x_n$$

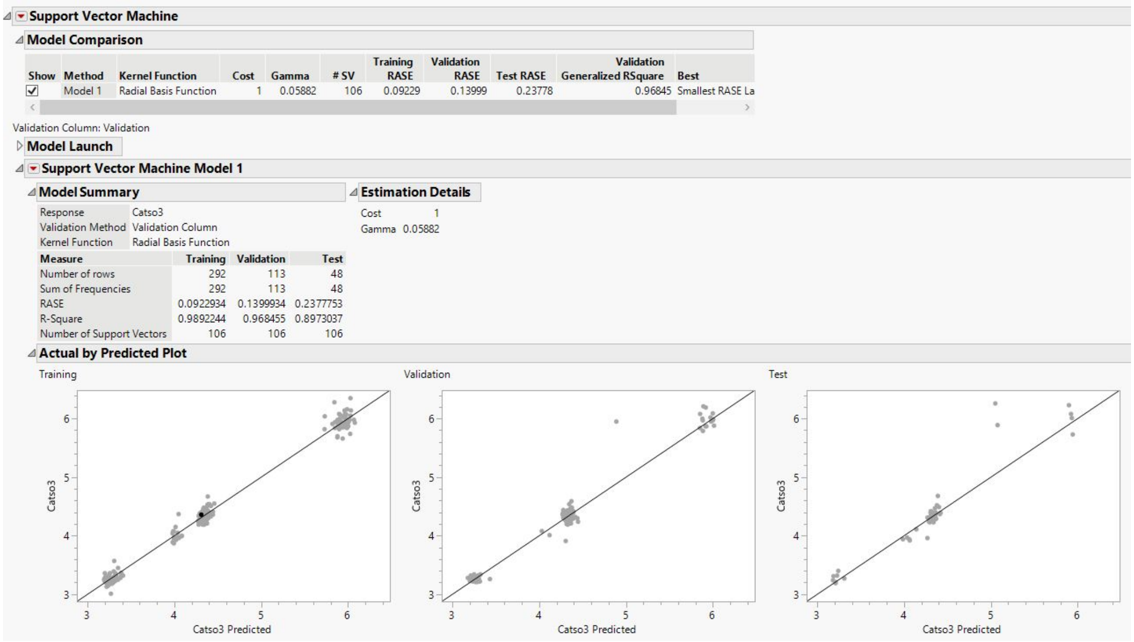
Constructing a Lagrangian function from the primal function by introducing nonnegative multipliers a_n and a_n^* for each observation x_n. This leads to the formula, where we minimize

$$L(a) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(a_i - a_i^*)(a_j - a_j^*)x_i'x_j + \varepsilon\sum_{i=1}^{N}(a_i + a_i^*) + \sum_{i=1}^{N}y_i(a_i^* - a_i)$$

subject to the constraint:

$$\sum_{n=1}^{N}(a_n - a_n^*) = 0 \text{ [10]}$$

As in the previous section, we run the model for SVM with selected features and all features separately. If we look at the Figure 3.14 and Figure 3.15 there is a few difference between two model. And again we can visually see that the predictions of selected features are showing better results. Also with the high R square statistics results from SVR model, we know that linear kernel function was the best choice to predict our process response variables. Model also lets us use profiler again, by GLR comparison we can also print the vector parameters to predict which features are more important to our prediction. Profiler in this case will be a summary of our vector. The RASE value is a special statistics value which has a close meaning to mean square error statistics. It represents the square root of the mean squared prediction error. In the results section it will help us to identify which model is better according to real life error reduction.



Support Vector Machine

Model Comparison

Show	Method	Kernel Function	Cost	Gamma	# SV	Training RASE	Validation RASE	Test RASE	Validation Generalized RSquare	Best
☑	Model 1	Radial Basis Function	1	0.05882	106	0.09229	0.13999	0.23778	0.96845	Smallest RASE La

Validation Column: Validation

Model Launch

Support Vector Machine Model 1

Model Summary

Response	Catso3
Validation Method	Validation Column
Kernel Function	Radial Basis Function

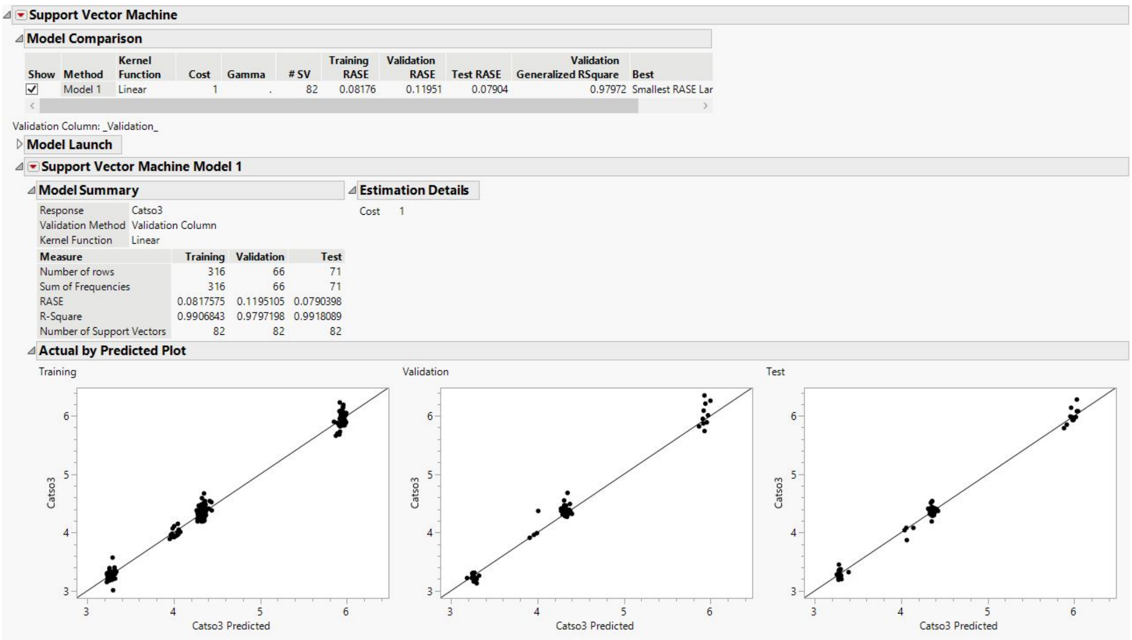
Measure	Training	Validation	Test
Number of rows	292	113	48
Sum of Frequencies	292	113	48
RASE	0.0922934	0.1399934	0.2377753
R-Square	0.9892244	0.968455	0.8973037
Number of Support Vectors	106	106	106

Estimation Details

Cost	1
Gamma	0.05882

Actual by Predicted Plot

Figure 3.14: SVM with all data



Support Vector Machine

Model Comparison

Show	Method	Kernel Function	Cost	Gamma	# SV	Training RASE	Validation RASE	Test RASE	Validation Generalized RSquare	Best
☑	Model 1	Linear	1	.	82	0.08176	0.11951	0.07904	0.97972	Smallest RASE Lar

Validation Column: _Validation_

Model Launch

Support Vector Machine Model 1

Model Summary

Response	Catso3
Validation Method	Validation Column
Kernel Function	Linear

Measure	Training	Validation	Test
Number of rows	316	66	71
Sum of Frequencies	316	66	71
RASE	0.0817575	0.1195105	0.0790398
R-Square	0.9906843	0.9797198	0.9918089
Number of Support Vectors	82	82	82

Estimation Details

Cost	1

Actual by Predicted Plot

Figure 3.15: SVM with selected features and adjusted validation data set

3.3 Bootstrap Forest

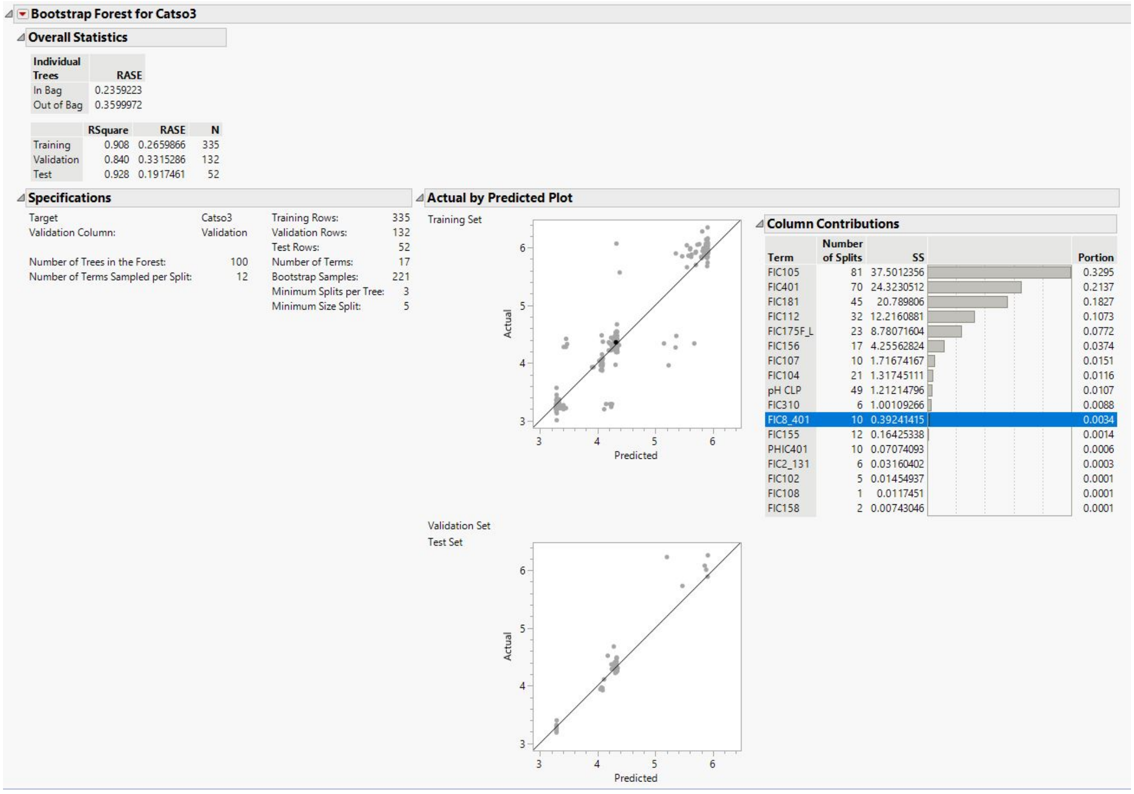
”To be able to explain Bootstrap Forest modelling. First we need to understand how decision trees work. Decision tree is a model that predicts the target value by learning simple decision rules which are inferred from the data features. The Decision Trees consist of a root node, internal also referred to as test nodes, and leaf nodes also called terminal or decision nodes. These nodes together assemble a directed rooted tree where the root node has no incoming edge. On the contrary, internal and leaf nodes have exactly one incoming edge. The crucial difference between internal and leaf nodes is that internal nodes also have outgoing edges. Another characteristic is that each test node split the instance space into two or more sub-spaces according to a discrete function of the input features. Deeper trees produce models that are fitted better and implement more complex rules. Generally, less complex decision trees are considered to be more comprehensible; moreover, the complexity of the tree has an essential effect on the model accuracy.” [11] In our model (continuous response value y), the Sum of Squares are reported. This change happens in the error sum-of-squares because of the split. A chosen candidate SS is:

$$SS_{test} = SS_{parent} \text{ - } (SS_{right} + SS_{left}) \text{ where SS in a node is just } s^2(n-1).$$

Bootstrapping is a statistical re-sampling technique that involves random sampling of a dataset with replacement. It is often used for quantifying the uncertainty related to model that is created. Random forest also called decision trees, is an assembly technique which separates the variables and gathers them in the same pools by tree like structures where every branch has rules for separation. In the regression trees, they start from the root of the tree and follow splits based on variable outcomes until a leaf node is reached and the result is given. The rules are basically asks the variables whether they are less,equal or greater than a certain number then through the yes-no answer gives the direction to the tree where the splitting happens.

The bootstrapping Random Forest algorithm combines ensemble learning method with the decision tree framework (in our case continues responses) to create multiple randomly drawn decision trees from the data, averaging the results to output a new result that often leads to strong prediction.

Bootstrap forest method usually works good with categorization data. Although we are dealing with continuous responses by fine tuning the decision trees we can have acceptable R square values. Adjusting the model parameters for the algorithm is a process itself. According to adjustment, meaning selecting tree depths, criterion and minimum sample splits, model could give really good responses. In this process we have decided to go with the suggested parameters of the model package. If we compare the model on the basis of all features vs selected features in Figure 3.16 and Figure 3.18 we say it is clear that model with selected features are predicted better then model with all features. Now the question is: What caused this improvement? Is it that the standard tree dept number and maximum tree size maybe more adjusted to the selected variable data set since it has less features? I believe that is not the case since decision trees are immune to number of feature sizes, meaning in their leaves they are already doing the splits and by that way choosing the right features and reducing the size of the features in its core automatically. Our Forest mainly focused on approximately four variables, meaning even our selected featured data set mainly too much for our model. However the difference between predicted value approximation mainly caused because of the validation data set. When we created the validation data set for selected featured model, what we included in the validation protocol is actually the category variable itself. We literally blended the categorical variable inside the data set, so when decision trees started to run training, validation and test sets separately



Bootstrap Forest for Catso3

Overall Statistics

Individual Trees	RASE
In Bag	0.2359223
Out of Bag	0.3599972

	RSquare	RASE	N
Training	0.908	0.2659866	335
Validation	0.840	0.3315286	132
Test	0.928	0.1917461	52

Specifications

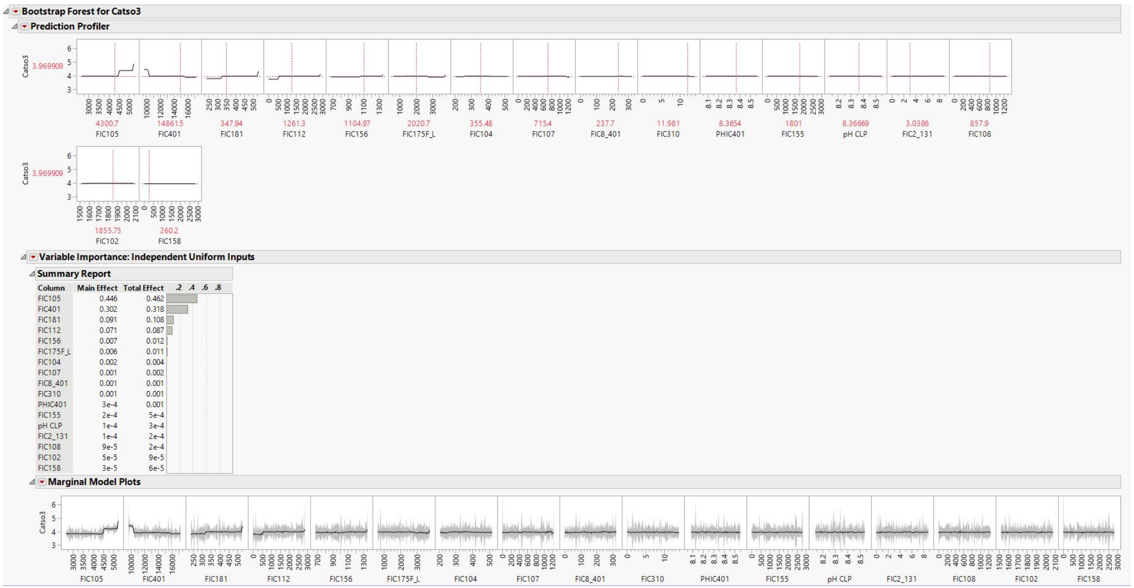
Target	Catso3	Training Rows:	335
Validation Column:	Validation	Validation Rows:	132
		Test Rows:	52
Number of Trees in the Forest:	100	Number of Terms:	17
Number of Terms Sampled per Split:	12	Bootstrap Samples:	221
		Minimum Splits per Tree:	3
		Minimum Size Split:	5

Actual by Predicted Plot

Column Contributions

Term	Number of Splits	SS	Portion
FIC105	81	37.5012356	0.3295
FIC401	70	24.3230512	0.2137
FIC181	45	20.789806	0.1827
FIC112	32	12.2160881	0.1073
FIC175F_L	23	8.78071604	0.0772
FIC156	17	4.25562824	0.0374
FIC107	10	1.71674167	0.0151
FIC104	21	1.31745111	0.0116
pH CLP	49	1.21214796	0.0107
FIC310	6	1.00109266	0.0088
FIC8_401	10	0.39241415	0.0034
FIC155	12	0.16425338	0.0014
PHIC401	10	0.07074093	0.0006
FIC2_131	6	0.03160402	0.0003
FIC102	5	0.01454937	0.0001
FIC108	1	0.0117451	0.0001
FIC158	2	0.00743046	0.0001

Figure 3.16: Bootstrap Forest with all features 1



Bootstrap Forest for Catso3

Prediction Profiler

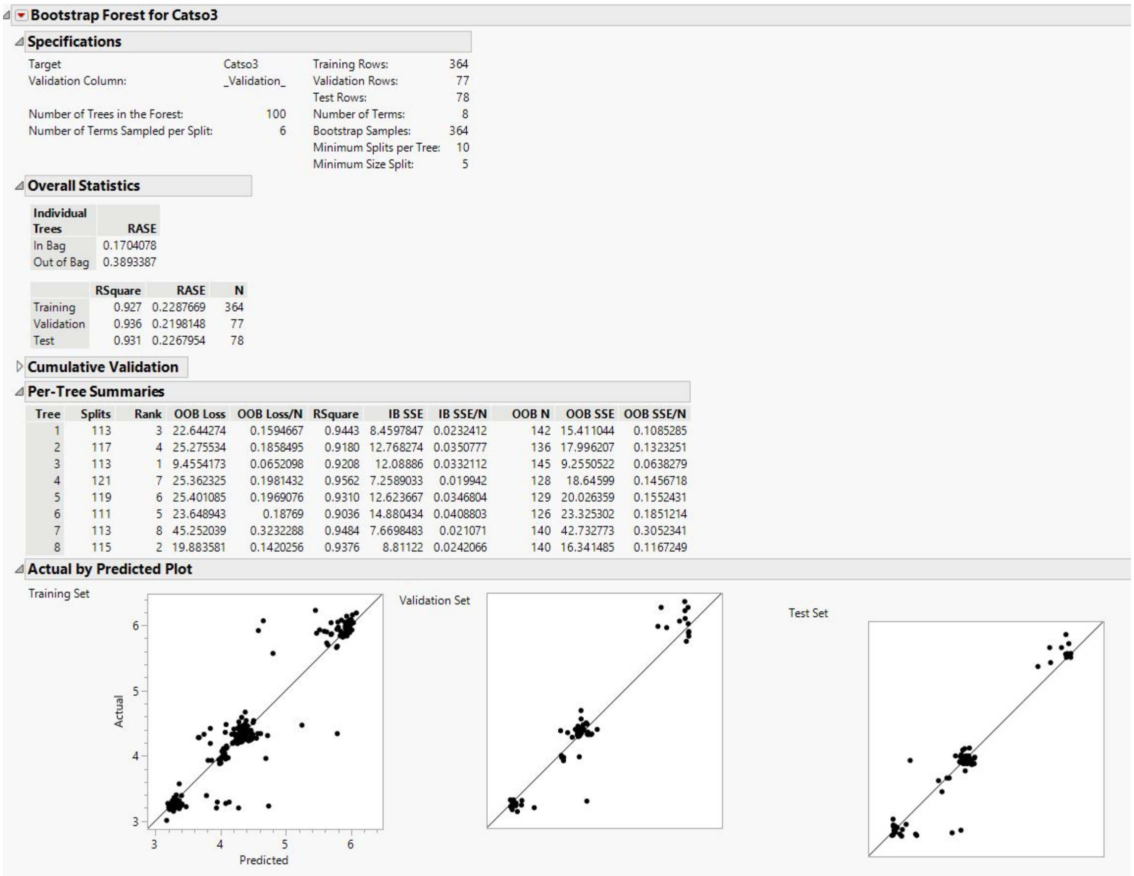
Variable Importance: Independent Uniform Inputs

Summary Report

Column	Main Effect	Total Effect
FIC105	0.446	0.462
FIC401	0.302	0.318
FIC181	0.091	0.108
FIC112	0.071	0.087
FIC156	0.007	0.012
FIC175F_L	0.006	0.011
FIC104	0.002	0.004
FIC107	0.001	0.002
FIC8_401	0.001	0.001
FIC310	0.001	0.001
PHIC401	3e-4	0.001
FIC155	2e-4	5e-4
pH CLP	1e-4	3e-4
FIC2_131	1e-4	2e-4
FIC108	9e-5	2e-4
FIC102	5e-5	9e-5
FIC158	3e-5	6e-5

Marginal Model Plots

Figure 3.17: Bootstrap Forest with all features 2



Bootstrap Forest for Catso3

Specifications

Target	Catso3	Training Rows:	364
Validation Column:	_Validation_	Validation Rows:	77
		Test Rows:	78
Number of Trees in the Forest:	100	Number of Terms:	8
Number of Terms Sampled per Split:	6	Bootstrap Samples:	364
		Minimum Splits per Tree:	10
		Minimum Size Split:	5

Overall Statistics

Individual Trees	RASE
In Bag	0.1704078
Out of Bag	0.3893387

	RSquare	RASE	N
Training	0.927	0.2287669	364
Validation	0.936	0.2198148	77
Test	0.931	0.2267954	78

Cumulative Validation

Per-Tree Summaries

Tree	Splits	Rank	OOB Loss	OOB Loss/N	RSquare	IB SSE	IB SSE/N	OOB N	OOB SSE	OOB SSE/N
1	113	3	22.644274	0.1594667	0.9443	8.4597847	0.0232412	142	15.411044	0.1085285
2	117	4	25.275534	0.1858495	0.9180	12.768274	0.0350777	136	17.996207	0.1323251
3	113	1	9.4554173	0.0652098	0.9208	12.08886	0.0332112	145	9.2550522	0.0638279
4	121	7	25.362325	0.1981432	0.9562	7.2589033	0.019942	128	18.64599	0.1456718
5	119	6	25.401085	0.1969076	0.9310	12.623667	0.0346804	129	20.026359	0.1552431
6	111	5	23.648943	0.18769	0.9036	14.880434	0.0408803	126	23.325302	0.1851214
7	113	8	45.252039	0.3232288	0.9484	7.6698483	0.021071	140	42.732773	0.3052341
8	115	2	19.883581	0.1420256	0.9376	8.81122	0.0242066	140	16.341485	0.1167249

Actual by Predicted Plot

Figure 3.18: Bootstrap Forest with selected features and adjusted validation data set

with bootstrap method, the R squares become much much more acceptable. In the side note that improvement proves that decision trees work much better with classification problems with categorical data sets.

In the results section we will provide the comparison of the model with more detail but with statistical point of view the model is acceptable and valid. Inclusion of the random forest in this study is to mainly show about the importance of validation data set adjustment, how it affected the results. That is also proves that even though we are reducing our features, according to our data set, we can blend some of the features into the consideration of the model while it does the validation.

3.4 Neural Network

Our last model is Neural Network. First I would like to start with summarizing of the model, and how it works. ”Neural networks are computing systems with interconnected nodes that work much like neurons in the human brain. Using algorithms, they can recognize hidden patterns and correlations in raw data, cluster and classify it, and – over time – continuously learn and improve. A simple neural network includes an input layer, an output (or target) layer and, in between, a hidden layer. The layers are connected via nodes, and these connections form a “network” – the neural network – of interconnected nodes. As the number of hidden layers within a neural network increases, deep neural networks are formed. Data is fed into a neural network through the input layer, which communicates to hidden layers. Processing takes place in the hidden layers through a system of weighted connections. Nodes in the hidden layer then combine data from the input layer with a set of coefficients and assigns appropriate weights to inputs. These input-weight products are then summed up. The sum is passed through a node’s activation function, which determines the extent that a signal must progress further through the network to affect the final output. Finally, the hidden layers link to the output layer – where the outputs are retrieved.” [12]

In our model we decided to not go into deep learning part since it becomes more complex and hard to explain as process vise. However we will use boosted neural networks. Boosting is the process of building a large additive neural network model by fitting a sequence of smaller models. Each of the smaller models is fit on the scaled residuals of the previous model. The models are combined to form the larger final model. The process uses validation to assess how many component models to fit, not exceeding the specified number of models. Boosting is often faster than fitting a single large model. However, the base model should be a 1 to 2 node single-layer model. The benefit of faster fitting can be lost if a large number of models is specified [13]. The model that we made has 1 node meaning one layer neural network. It decreased the time of training by ten fold. Also for the activation we have chosen Tanh:

$$TanH = f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \qquad \text{Figure 3.19}$$

Also we believe that all the features are contributing to prediction. That is why we used Squared penalization method and see all the predictors are contributing to the predictive ability of the model.

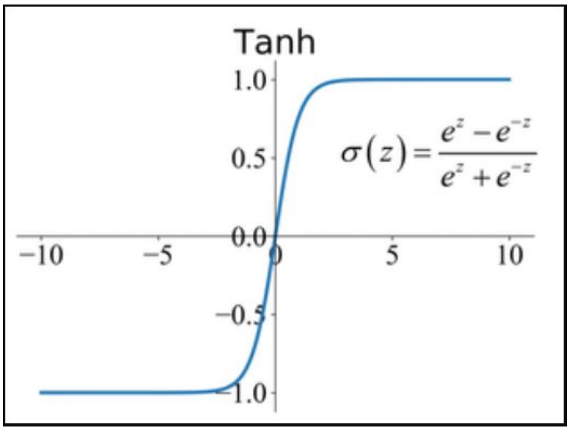


Figure 3.19: Neural Network with selected features

$$\text{Learning (over simplified): } \eta p(w_i) = \eta \sum w_i^2$$

The learning is ηp(w_i), where η is the learning parameter, and p( ) is a function of the parameter estimates, called the learning function. Validation is used to find the optimal value of the learning parameter.For the model the equation of such a single-layer neural network in terms of the individual components:

$$y_k = \sum_{i=0}^{d} w_{ki} x_i$$

We can summarise the terminology we have just introduced:

- Input vector $x = (x_0, x_1, ..., x_d)^T$
- Output vector $y = (y_1, ..., y_k)^T$
- Weight matrix W: w_{ki} is the weight from input x_i to output y_k

Now we have to train the network. The network is trained using a training set that contains N input/output pairs (x_n, t_n) : 1≤n≤N, where $tn = (t_{n1}, ..., t_{nK})$ is the target output vector for input vector x_n. The error function should measure the distance of the output vectors y_n from the corresponding target vectors t_n for all n. A natural way to do this is by taking the (squared) Euclidean distance, and we define the sum-of-squares error function which computes the sum of squared Euclidean distances between t_n and y_n for all members of the training set 1≤n≤N. In matrix form we can write:

$$E(W) = \frac{1}{2} \sum_{n=1}^{N} \sum_{k=1}^{K} (y_{nk} - t_{nk})^2$$

Now we have to include the gradient descent to our algorithm. Gradient descent is an important optimisation technique, that may be used whenever it is possible to compute the derivatives of the error function with respect to the parameter to be optimised. For single layer neural networks, this means taking the derivative of the error function E(W) with respect to the weight matrix W. The idea of gradient descent is that to minimise an error function with respect to the parameters, we want to take small steps in a downhill direction. We take small steps because the gradient is not uniform, and if we take too big a step we may end up going uphill again! When considering this form of optimisation, we are considering another multidimensional space, weight space. This is a $K(d+1)$ dimension space, and a specific weight matrix W corresponds to a point in weight space. The error function evaluates the error value for a point in weight space (given the training set). Descending in weight space means adjusting the weight matrix W by moving a small direction down the gradient, which is the direction along which E decreases most rapidly. This means adjusting the weight factor in the direction of $-\nabla$WE, or adjusting each weight w_{ki} by adding a factor $\eta E/w_{ki}$, where η is a small constant called the learning rate. If we write the value of a weight at iteration τ as w_{ki}^{τ}, then its updated value is given by:

$$w_{ki}^{\tau+1} = w_{ki}^{\tau} - \eta \frac{\partial E}{\partial w_{ki}} \text{ [14]}$$

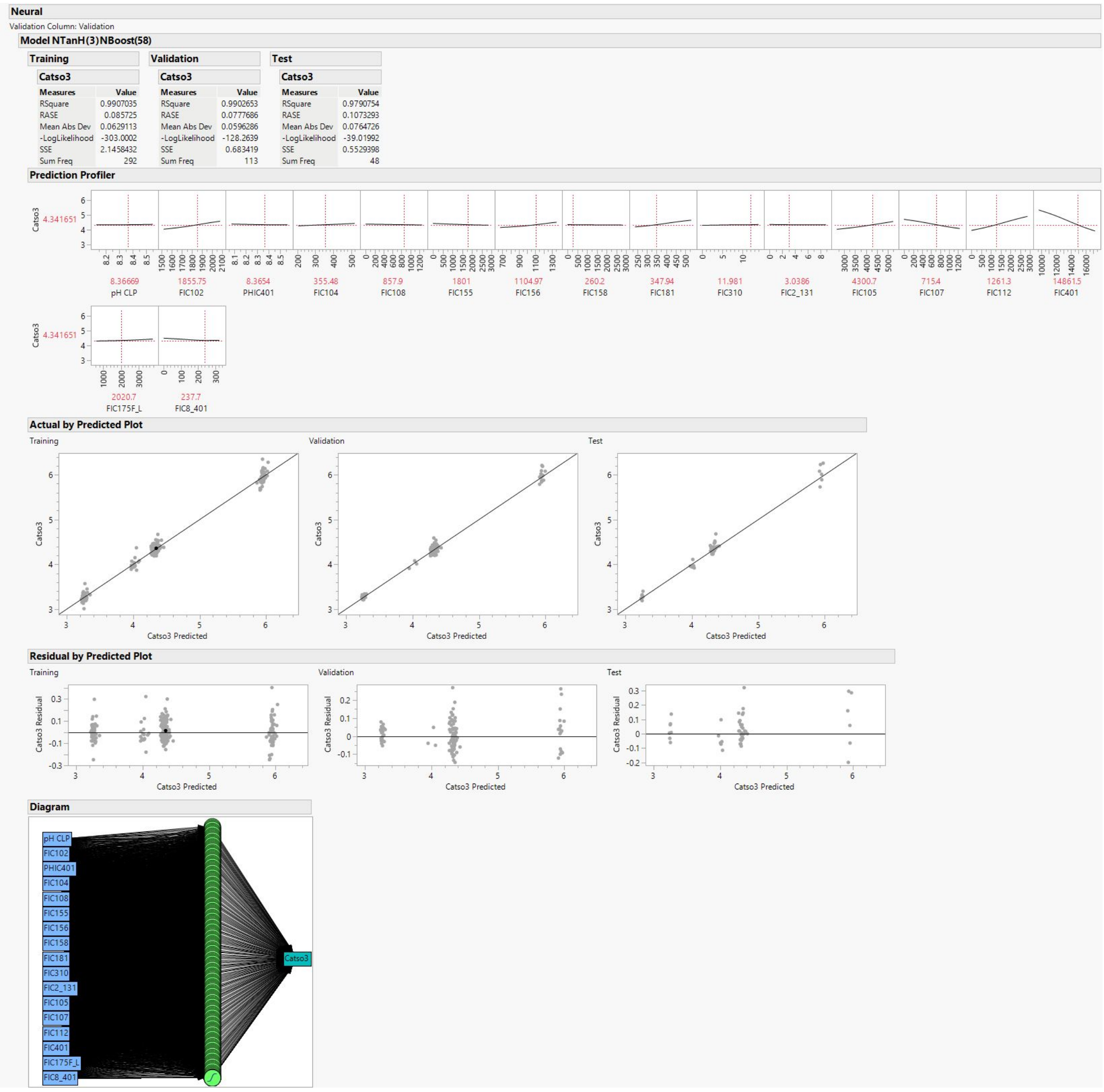
In the end our model will sum up to below, where the g() function will include our boosting and activation.

$$y_k(x_i) = g(\sum_{i=0}^{d} w_{ki}x_i)$$

As previous cases we also did two neural network models Figure 3.20 Figure 3.21 for selected features and one with the adjusted validation data set with selected features Figure 3.22. The diagrams that are presented all the figures in this section looks complicated since that is the part boosting applied. For all features model Figure 3.20 the number of boosting is 58. For the selected features model Figure 3.21 number of boosting is reduced to 53 and finally for the adjusted validation data set with selected features model that number reduced to 49. Every time we move forward to adjust the parameters in our favors the required number of boosting decreases. That is another way of saying that we are going into right direction. The network diagram for selected features and adjusted validation data set model is Figure 3.23.

It is important to notice that the profiler section graph's slopes easily tells us which features play the active role in our model. If we look at the slopes of the profilers in all the models the surfactants are playing big role in the prediction of response values. One other topic that we see on the results are residuals and prediction graphs. The prediction graphs are showing really close results to actual responses and residuals can be counted between +0.3 to -0.2 which is considerably good even for process engineering perspective.

For neural network session, we just would like to compare our neural networks between each other. If we look at the R square results and all the statistical parameters that the models represent, all of them are significantly valid. What we can do is to separate them by little differences they have and choose one of them to consider in results section. There are literally in third degree decimal differences between them. For consideration it is obvious that test and validation data sets will be effective to do the comparison. What it is important for us as process point of view is the test set. Because the test set is the values which are separated completely from the adjustment and creation of the models. Taking into consideration of test sets performance of R square and Root average Square



Neural

Validation Column: Validation

Model NTanH(3)NBoost(58)

Training

Catso3

Measures	Value
RSquare	0.9907035
RASE	0.085725
Mean Abs Dev	0.0629113
-LogLikelihood	-303.0002
SSE	2.1458432
Sum Freq	292

Validation

Catso3

Measures	Value
RSquare	0.9902653
RASE	0.0777686
Mean Abs Dev	0.0596286
-LogLikelihood	-128.2639
SSE	0.683419
Sum Freq	113

Test

Catso3

Measures	Value
RSquare	0.9790754
RASE	0.1073293
Mean Abs Dev	0.0764726
-LogLikelihood	-39.01992
SSE	0.5529398
Sum Freq	48

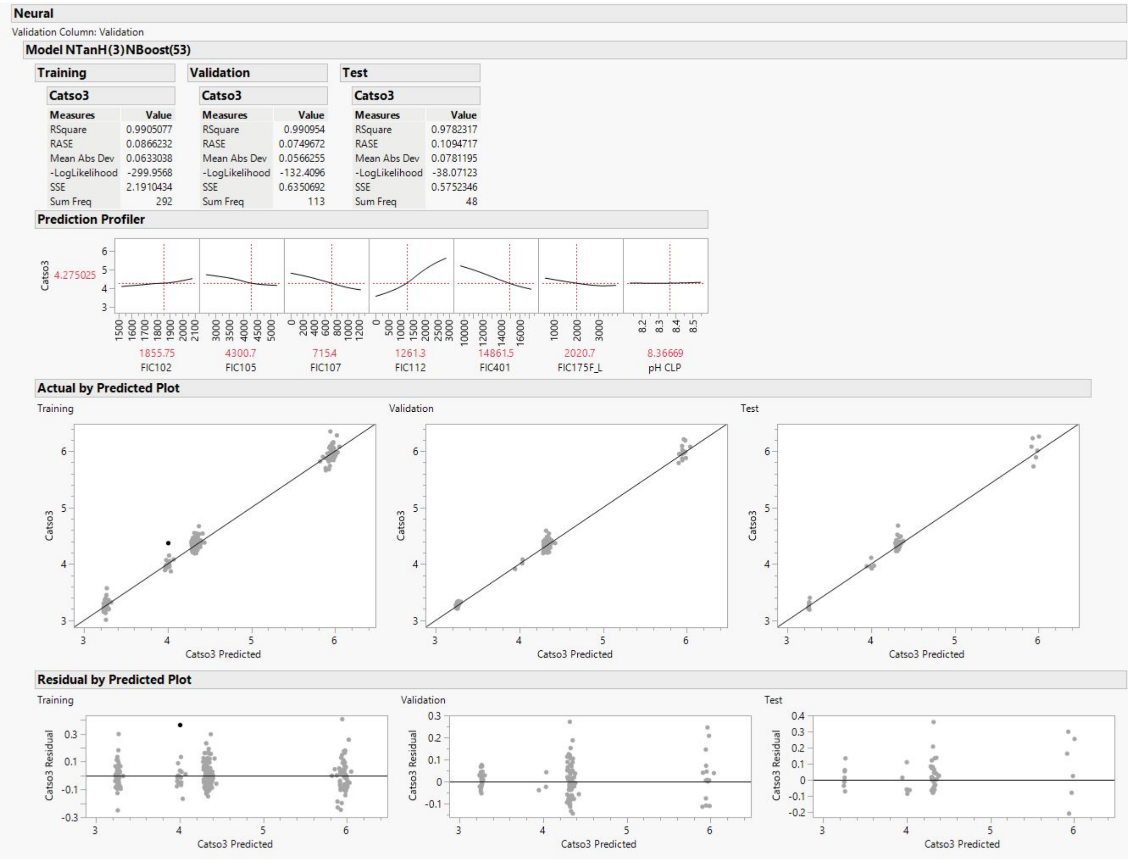
Prediction Profiler

Actual by Predicted Plot

Residual by Predicted Plot

Diagram

Figure 3.20: Neural Network with all the features



Neural

Validation Column: Validation

Model NTanH(3)NBoost(53)

Training

Catso3

Measures	Value
RSquare	0.9905077
RASE	0.0866232
Mean Abs Dev	0.0633038
-LogLikelihood	-299.9568
SSE	2.1910434
Sum Freq	292

Validation

Catso3

Measures	Value
RSquare	0.990954
RASE	0.0749672
Mean Abs Dev	0.0566255
-LogLikelihood	-132.4096
SSE	0.6350692
Sum Freq	113

Test

Catso3

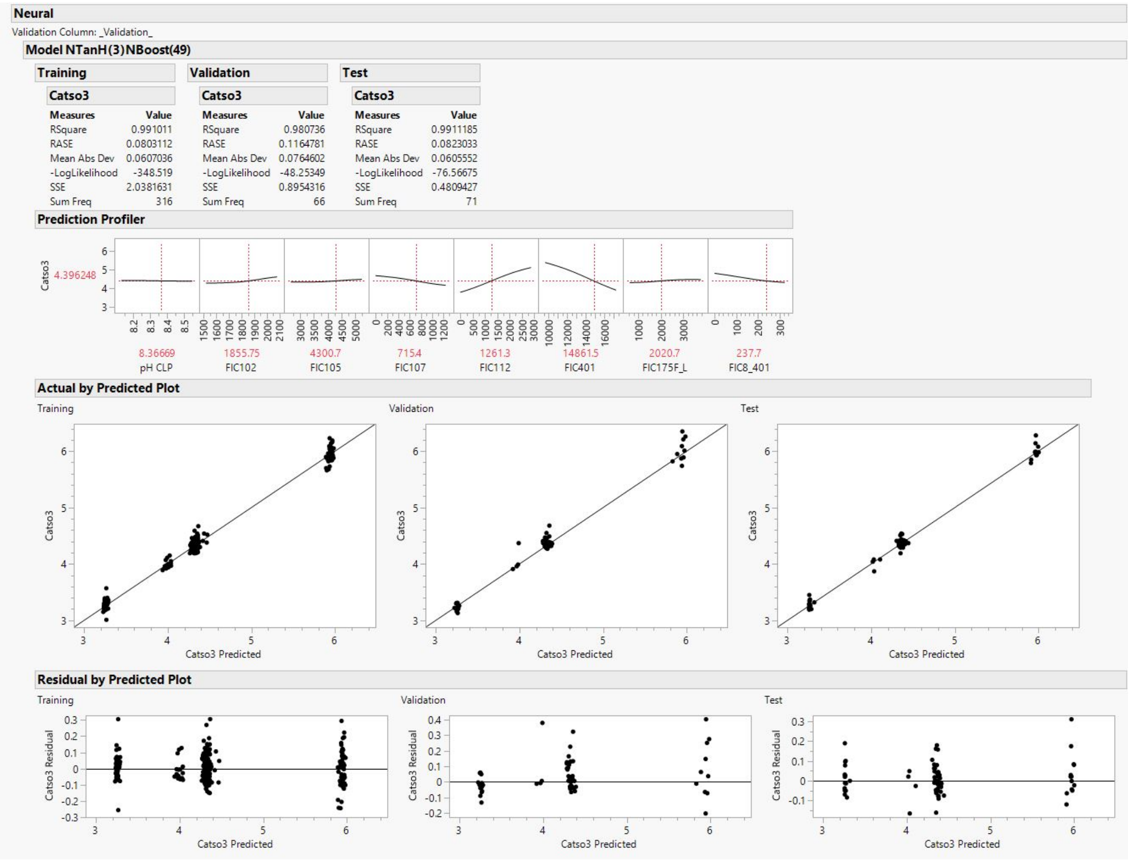
Measures	Value
RSquare	0.9782317
RASE	0.1094717
Mean Abs Dev	0.0781195
-LogLikelihood	-38.07123
SSE	0.5752346
Sum Freq	48

Prediction Profiler

Actual by Predicted Plot

Residual by Predicted Plot

Figure 3.21: Neural Network with selected features



Neural

Validation Column: _Validation_

Model NTanH(3)NBoost(49)

Training — Catso3

Measures	Value
RSquare	0.991011
RASE	0.0803112
Mean Abs Dev	0.0607036
-LogLikelihood	-348.519
SSE	2.0381631
Sum Freq	316

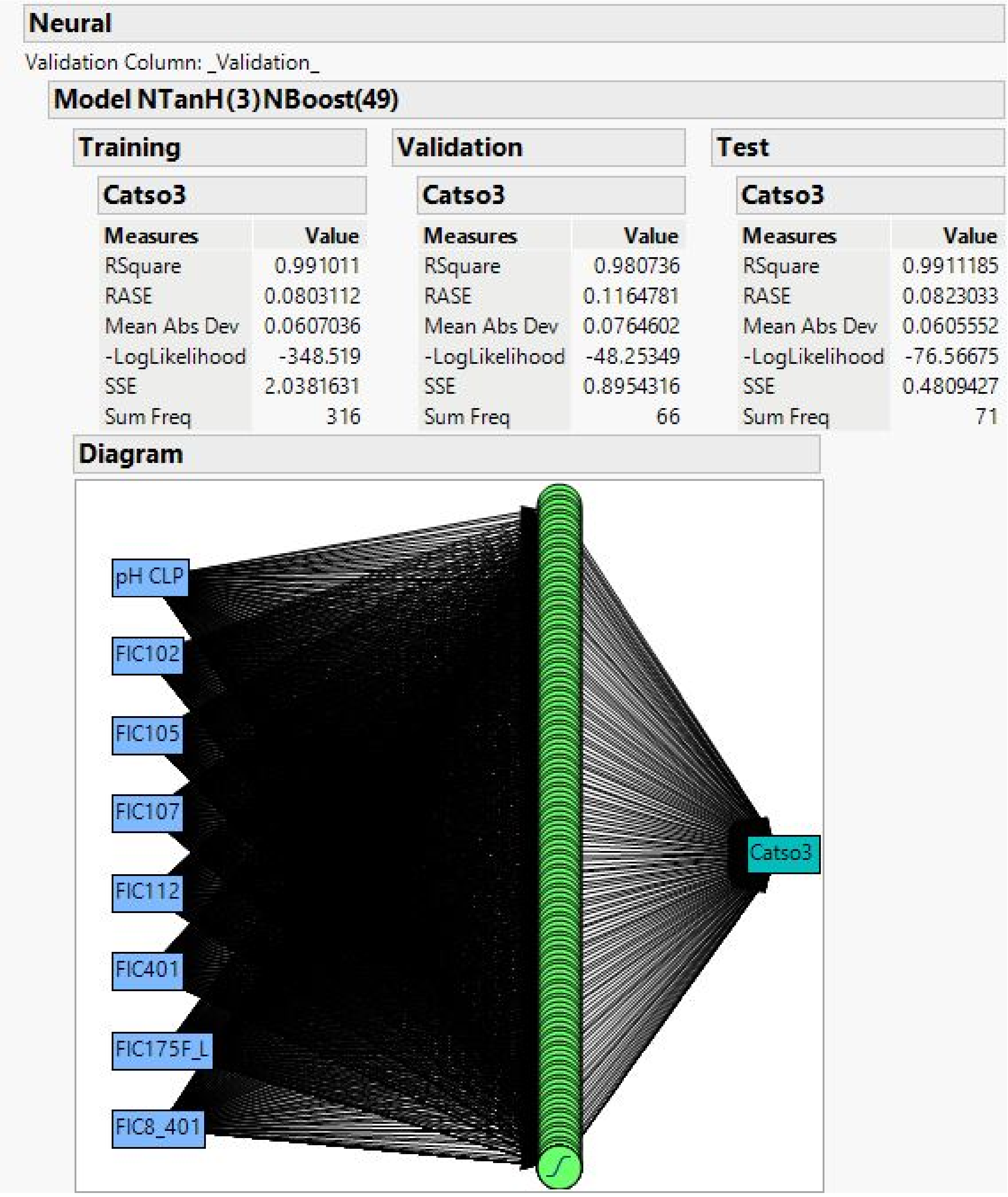
Validation — Catso3

Measures	Value
RSquare	0.980736
RASE	0.1164781
Mean Abs Dev	0.0764602
-LogLikelihood	-48.25349
SSE	0.8954316
Sum Freq	66

Test — Catso3

Measures	Value
RSquare	0.9911185
RASE	0.0823033
Mean Abs Dev	0.0605552
-LogLikelihood	-76.56675
SSE	0.4809427
Sum Freq	71

Figure 3.22: Neural Network with selected features and adjusted validation data set



Neural

Validation Column: _Validation_

Model NTanH(3)NBoost(49)

Training

Catso3

Measures	Value
RSquare	0.991011
RASE	0.0803112
Mean Abs Dev	0.0607036
-LogLikelihood	-348.519
SSE	2.0381631
Sum Freq	316

Validation

Catso3

Measures	Value
RSquare	0.980736
RASE	0.1164781
Mean Abs Dev	0.0764602
-LogLikelihood	-48.25349
SSE	0.8954316
Sum Freq	66

Test

Catso3

Measures	Value
RSquare	0.9911185
RASE	0.0823033
Mean Abs Dev	0.0605552
-LogLikelihood	-76.56675
SSE	0.4809427
Sum Freq	71

Diagram

Figure 3.23: Neural Network Diagram

Error (RASE), we can easily say that the model with selected features and adjusted validation data set came first in our list. It should be noted that all the models are exceptionally good but we just want to reduce to comparison to models based results by eliminating same models that behave less good. We will also explain why we are looking at RASE values more than R Square values in the results section.

4
Results

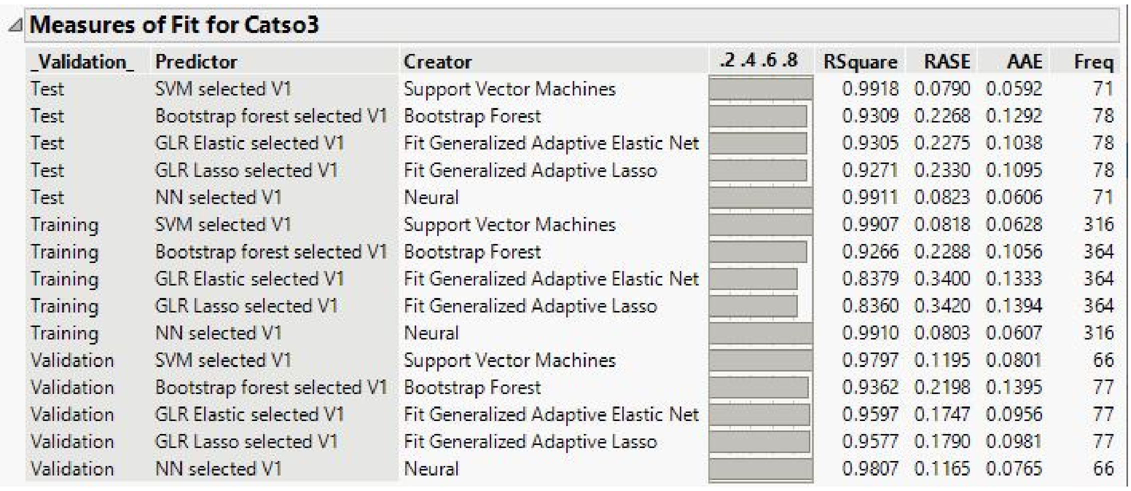
The main focus of this study is to come up with a model which can predict a response variable, in our case is the CatSO3 values, with a minimum error in a continuous liquid process manufacturing. Our liquid is called whitebase and all the features that we use to predict the response variable are given by process engineers. It is a complete closed process leaving little to no error. That is why it is expected to deliver such model that can predict CatSO3 values with little to no error.

In the research that we have completed, we have applied many statistical techniques to:

- Reduce error of prediction
- Reduce feature size of the data set
- Reduce the data transformation
- increase the interpretability

With the techniques we have used, we were able to reduce the feature size to 9 instead of 21. To reduce the feature size first we plotted the correlation matrix of features. As we have seen in the previous section, we have removed the correlated features. To do this we have considered the correlation of features between each others and their correlation to the response variable. We have picked the ones which has high correlation with response variable and eliminated the others who are still correlated to this variable but less correlated to response variables. In any case to make sure that we have chosen the correct features, we also did modelling 2 times ones with the selected features and ones with the all features. In every model we run, selected features gave us better results. So in the result part we only included the selected featured models of each type.

Before going to the results section we also looked into our validation data set. There was a feature called brand in our data sets and we were not able to use it as a feature to create the model, since they tend to change by production decision. However while creating our validation data set we blend in the the brand feature into separation of the data via validation, training and test. Data set we created proportionally divided into sets by considering the



Measures of Fit for Catso3

Validation	Predictor	Creator	.2 .4 .6 .8	RSquare	RASE	AAE	Freq
Test	SVM selected V1	Support Vector Machines		0.9918	0.0790	0.0592	71
Test	Bootstrap forest selected V1	Bootstrap Forest		0.9309	0.2268	0.1292	78
Test	GLR Elastic selected V1	Fit Generalized Adaptive Elastic Net		0.9305	0.2275	0.1038	78
Test	GLR Lasso selected V1	Fit Generalized Adaptive Lasso		0.9271	0.2330	0.1095	78
Test	NN selected V1	Neural		0.9911	0.0823	0.0606	71
Training	SVM selected V1	Support Vector Machines		0.9907	0.0818	0.0628	316
Training	Bootstrap forest selected V1	Bootstrap Forest		0.9266	0.2288	0.1056	364
Training	GLR Elastic selected V1	Fit Generalized Adaptive Elastic Net		0.8379	0.3400	0.1333	364
Training	GLR Lasso selected V1	Fit Generalized Adaptive Lasso		0.8360	0.3420	0.1394	364
Training	NN selected V1	Neural		0.9910	0.0803	0.0607	316
Validation	SVM selected V1	Support Vector Machines		0.9797	0.1195	0.0801	66
Validation	Bootstrap forest selected V1	Bootstrap Forest		0.9362	0.2198	0.1395	77
Validation	GLR Elastic selected V1	Fit Generalized Adaptive Elastic Net		0.9597	0.1747	0.0956	77
Validation	GLR Lasso selected V1	Fit Generalized Adaptive Lasso		0.9577	0.1790	0.0981	77
Validation	NN selected V1	Neural		0.9807	0.1165	0.0765	66

Figure 4.1: Model Comparison

brand. So in that case we eliminated the possibility of bias through brand. Each brand production run now has a proportionally equal saying into decision of the model parameters. We also proved that by running the model into two different validation data sets, one with the adjusted validation data set and one without. The results were always in favour of the one with the adjusted validation data set.

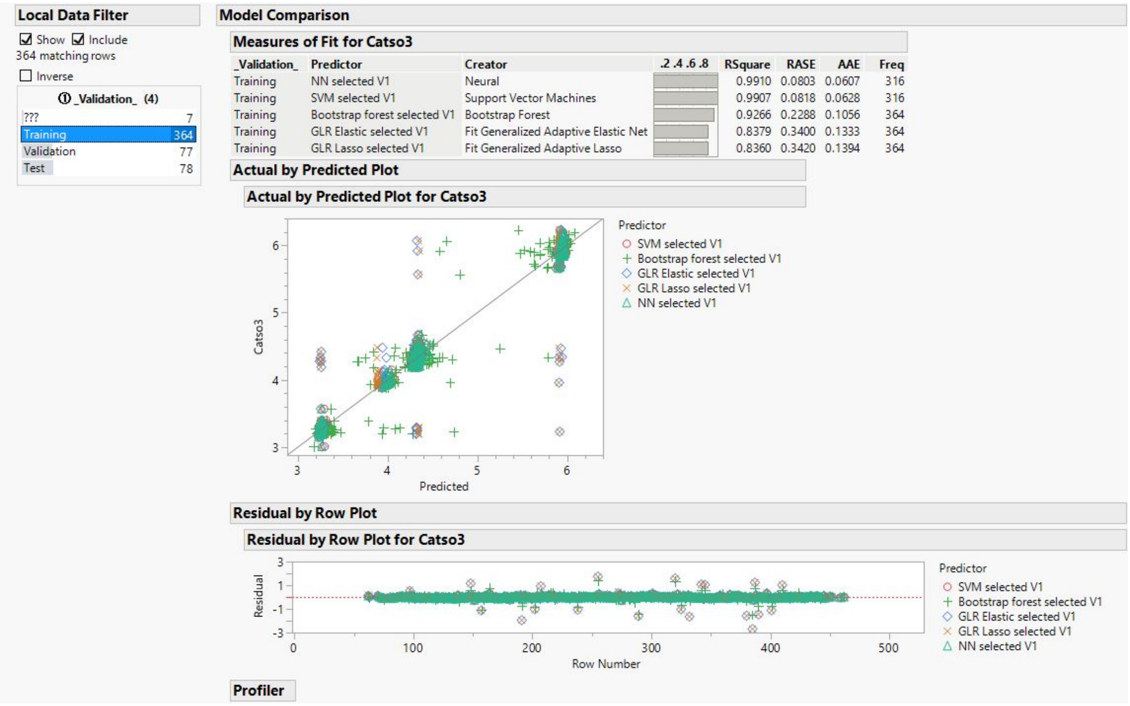
After receiving the result, we also used statistical techniques to evaluate and compare each model individually and cross comparatively. In the models section of this study, first we compared the models of each type separately between their own types. Now in this section we will compare them between types via selected models from each category. To do this we will look into their prediction graphs and their statistical results. The statistical results will include RASE and R square values.

Figure 4.1 is our summarizing table of each model divided by Test, Training and Validation dataset results. Figure 4.2,Figure 4.3 and Figure 4.4 shows us the distribution maps of the predicted values via models versus the real response values.

All the statistics that is presented by comparison tables and prediction plots for each model has really close values. The R squares for all of the model are above %97 and RASE values are below 0.4. All the AAE values are below 0.14. These are all good results according to a data scientist without the knowledge of the process. But before including the process point of view lets look into results in detail. Now I will compare the models in the category of validation data set separation. The validation data set is divided into %75 for training %15 for validation and %15 for test.

First lets look at the training data set results. For every aspect of statistics NN and SVM is taking the lead. They bot h have the highest R square, and lowest RASE and AAE results which is good. We started with the training data set because it includes the most number of observations. So in the overall comparison, from the trainig point of view we would select NN and SVM to be our model.

Second lets look at the validation part of the data set. For GLR and BT R square values are considerably increased with the addition RASE and AAE values decreased. Decrease might have caused by reduction of observation size but in any case SVM and NN is still taking the lead part. SVM and NN is competing in degree three



Local Data Filter

☑ Show ☑ Include

364 matching rows

☐ Inverse

① _Validation_ (4)	
???	7
Training	364
Validation	77
Test	78

Model Comparison

Measures of Fit for Catso3

Validation	Predictor	Creator	.2 .4 .6 .8	RSquare	RASE	AAE	Freq
Training	NN selected V1	Neural		0.9910	0.0803	0.0607	316
Training	SVM selected V1	Support Vector Machines		0.9907	0.0818	0.0628	316
Training	Bootstrap forest selected V1	Bootstrap Forest		0.9266	0.2288	0.1056	364
Training	GLR Elastic selected V1	Fit Generalized Adaptive Elastic Net		0.8379	0.3400	0.1333	364
Training	GLR Lasso selected V1	Fit Generalized Adaptive Lasso		0.8360	0.3420	0.1394	364

Actual by Predicted Plot

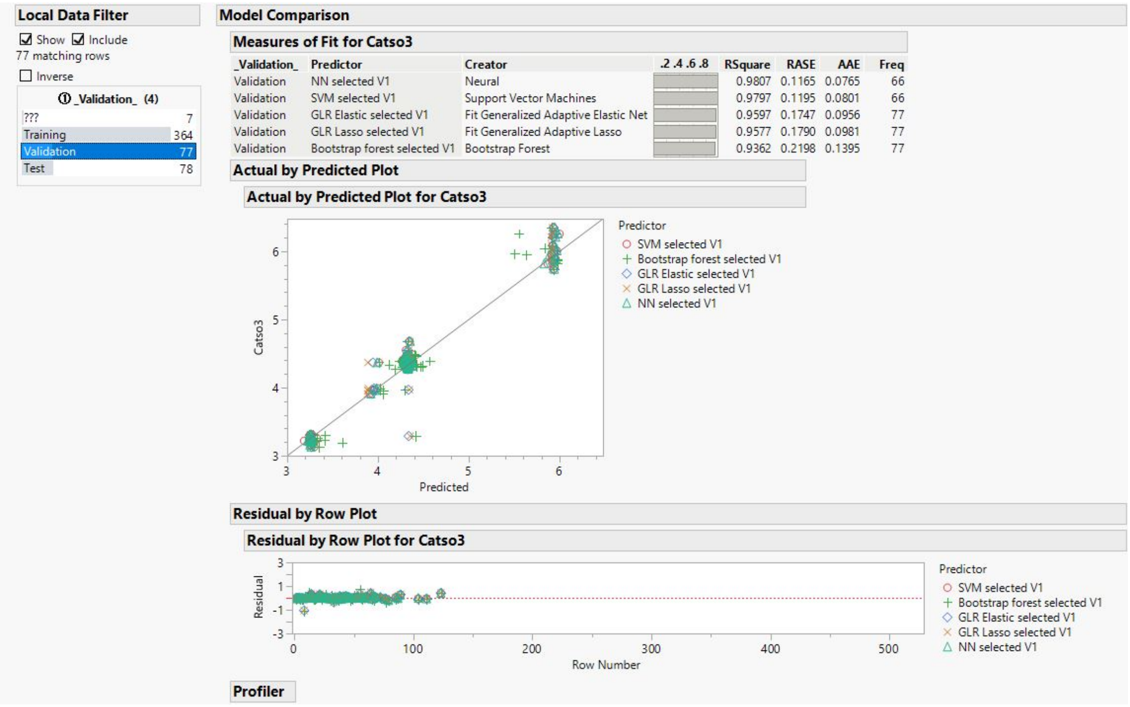
Actual by Predicted Plot for Catso3

Residual by Row Plot

Residual by Row Plot for Catso3

Profiler

Figure 4.2: Model Comparison Training results



Local Data Filter

☑ Show ☑ Include

77 matching rows

☐ Inverse

① _Validation_ (4)	
???	7
Training	364
Validation	77
Test	78

Model Comparison

Measures of Fit for Catso3

Validation	Predictor	Creator	.2 .4 .6 .8	RSquare	RASE	AAE	Freq
Validation	NN selected V1	Neural		0.9807	0.1165	0.0765	66
Validation	SVM selected V1	Support Vector Machines		0.9797	0.1195	0.0801	66
Validation	GLR Elastic selected V1	Fit Generalized Adaptive Elastic Net		0.9597	0.1747	0.0956	77
Validation	GLR Lasso selected V1	Fit Generalized Adaptive Lasso		0.9577	0.1790	0.0981	77
Validation	Bootstrap forest selected V1	Bootstrap Forest		0.9362	0.2198	0.1395	77

Actual by Predicted Plot

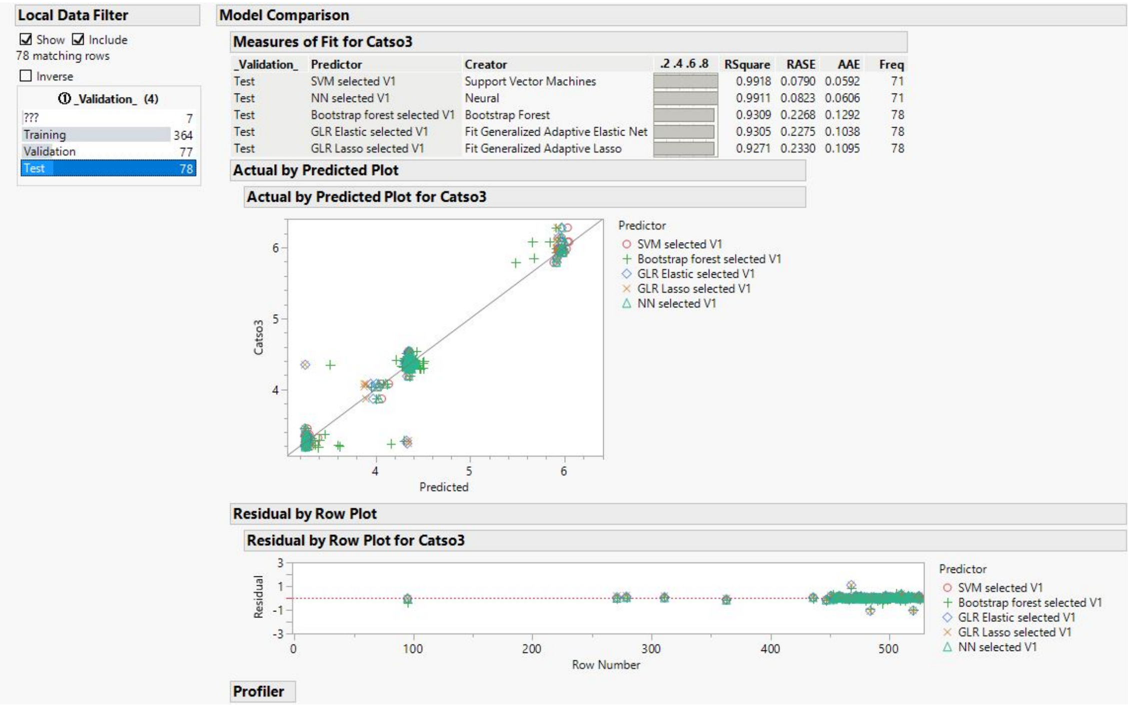
Actual by Predicted Plot for Catso3

Residual by Row Plot

Residual by Row Plot for Catso3

Profiler

Figure 4.3: Model Comparison Validation results



Local Data Filter

☑ Show ☑ Include
78 matching rows
□ Inverse

① _Validation_ (4)

???	7
Training	364
Validation	77
Test	78

Model Comparison

Measures of Fit for Catso3

Validation	Predictor	Creator	.2 .4 .6 .8	RSquare	RASE	AAE	Freq
Test	SVM selected V1	Support Vector Machines		0.9918	0.0790	0.0592	71
Test	NN selected V1	Neural		0.9911	0.0823	0.0606	71
Test	Bootstrap forest selected V1	Bootstrap Forest		0.9309	0.2268	0.1292	78
Test	GLR Elastic selected V1	Fit Generalized Adaptive Elastic Net		0.9305	0.2275	0.1038	78
Test	GLR Lasso selected V1	Fit Generalized Adaptive Lasso		0.9271	0.2330	0.1095	78

Actual by Predicted Plot

Actual by Predicted Plot for Catso3

Residual by Row Plot

Residual by Row Plot for Catso3

Profiler

Figure 4.4: Model Comparison Test results

decimals. Again in the statistical point of view all of them are decisive but to choose which one to be selected, SVM and NN will take the lead. In the third decimal after zero NN is better than SVM according to R square and lower according to RASE and AAE values.

Now lets look at the test part of the validation data set. Again for starters all the models look like they are performing great. GLR and BF is reduced in the third decimals for every statistical figure. GLR R square results decreased more than BF compared to validation part. Since it is the testing part, it is important that the changes that we see here actually express if there was overfitting in the previous sections. So we might come to and understanding by looking GLR model, there was a little overfitting to the model. Again NN and SVM takes the lead according to R square, RASE and AAE results. To sum up NN and SVM would be our primary choice if it comes to the selection between these 5 models. Now the important question rises since we would like to choose the best one. Which model is better for our need: NN or SWM?

To be able to decide which model is better between NN or SWM, we have exhausted our statistical parameters point of view. Now we would like to look at the selection from a process engineer point of view. Even though both models are near perfect situation we want to minimize our error. Prediction margins are a huge topic for continuous liquid production processes. The processes are considered effect proof and all the slight changes of measurements gives the process engineers an idea about how the process is going. There is little to no tolerance between measurements. That is being said, if we have a super model which predicts 99 values perfectly over 100 values and 1 value really terrible that model could not be used in the process. That 1 value would actually create an alarm and has to be investigated by process owners. It will gives us the information about either there is a measurement reader malfunction or there is an unidentified material which should not be in the closed circle in

Dataset	Model	R square	RASE	AAE
Test	SVM selected V1	0.9918	0.0790	0.0592
Test	NN selected V1	0.9911	0.0823	0.0606

Table 4.1: Closer look into NN and SVM model

the first place. Both of them will create a problem. First one will tell us that validation protocol of the measurement reader has to be changed or the reader itself has to be changed. The second one is a quality problem which has no toleration. The product has to be eliminated and reproduced. In both scenarios since the model will be okay with the results, it will omit that one terrible predicted value.

To avoid a single terrible observation prediction, in the results we have to look into the RASE and AAE results. Also we have to do it in the test validation set. Test values are the ones that have no effect on the model what so ever and that is why it is important to consider the test set overall. While the test set has a certain job in our model, it also has a side job to test the processes integrity. Lets look at the Table 4.1.

Our goal in this situation will be to reduce the error as much as possible according to a new data set which has no effect on model's creation. So in that case RASE and AAE play a really important role for our selection. By looking at Table 4.1, because of the reasoning of a process engineer and a data scientist we would choose SVM over NN even though error reduction is in the third decimal point of RASE value. So it is fair to say that our SVM model is the winner.

In the overall process, eventually there will be a discussion about how to improve our model. By the look of the results it seems like no improvements are needed but this is a continuous production process. Eventually the brands will change and so the level of ingredients. Instead of levelling between third decimal point, what could be done is to increase our sample size by ten fold. Increasing our sample size would increase the space between statistical values. Also including brand category we would increase the quality of statistical results. Both of these solutions look like feasible in the data scientist point of view but they are really expensive in the process engineer point of view. Every laboratory analysis, every retraining of the model costs capital. Introduction of a new brand will create a retraining cost to our model and these 529 observations in real life is already a years of work. That is why the study is limited to this data set. What could be done is to connect this model into data pipeline and continue to updating the model in the coming years. There could be two model which would predict only with this golden data set and the other would update itself with incoming data. After the results of updating model surpasses the eventual used stable model, we could put the better model into the production line and create another model which would continue with updating data set while the model which is put on the production stop updating.

Overall our selection from the model comparison would be Support Vector Regressor in terms of data scientist perspective and process engineer perspective. All the improvement could be done to better the model is out of scope of this research.

5 Conclusion

Machine learning (ML) is a field of study in artificial intelligence concerned with the development and study of statistical algorithms that can effectively generalize and thus perform tasks without explicit instructions [15]. Data science is a ”concept to unify statistics, data analysis, informatics, and their related methods” to ”understand and analyze actual phenomena” with data. It uses techniques and theories drawn from many fields within the context of mathematics, statistics, computer science, information science, and domain knowledge [16]. In this research we have combined data science knowledge and process engineering knowledge to come up with a machine learning method which will help the process improve its foundation. The start of the research initialized by the idea of cost reduction. The cost of laboratory analysis would be reduced by simply replacement of the manual analysis with a predictive machine learning system. So that laboratory would focus on different matters at hand.

The research topic was to predict a certain surfactant value (in our case is the CatSO3) inside the whitebase (which is the product) without needing the actual laboratory analysis. The data which has been used for the research are gathered via sensors of the product line and the laboratory analysis results. Our response variable was the laboratory results, but all the other values are gathered via sensors inside the production line. The production line is a closed circuit with a lot of sensors leaving no error to the process flow. The plant is located in Pomezia, Italy and owned by Procter&Gamble company. It is a plant that produces high density liquid and information share about the plant specifics, ingredients that have been used and process itself are strictly restricted. Due to non disclosure agreement the variable names have been coded into letters and numbers. All the overall information shared by this research is also available to public.

In the process side the importance of the ingredients have been explained and the rest of the research included data science project. In the data science part of the research different machine learning methods have been tested via given data sets. The data set have been cleaned and shaped with the help of explanatory data analysis techniques. Even before continuing the model creation, the data set also adjusted on the accordance of validation. With the help of production process knowledge the data have been separated to training, validation and test data sets. After that model creation of part of the research has been initiated. Machine learning models that have been created are focused on predictive regression analysis. These models were GLR-L, GLR-E, Rf, NN and SWM. All the models that have been used are summarized in the models section. In the same section it is also decided that which model of the same type will be used according to data set which has been used to create them.

In the results section firstly the models are compared via the statistical information they provide. In the statistical knowledge point of view all type of the model were valid and all them gave us a significant results. Among the model SWM and NN take the lead to be chosen into detailed consideration. These two models are very effective of predicting CatSO3 levels in the whitebase production. However we continued to compare them via process engineering criteria which was the reduce the error as much as possible, by doing so we take into account the test set of the validation process. The reason to choose the test set comparison is because test set have not been considered in the creation of the model and is a great stand point even for process point of view. Also to take into account every R square statistics of each model were racing in the third decimal point of the result. By accepting process knowledge, we went for the model with the lowest error of prediction. The lowest error for the models are decided by RASE and AAE statistics. After the comparison we have decided that SVM model is our best option to use for prediction of CatSO3 level in whitebase production.

To be able to improve the model what we could do is stated previously in the results section. The choices of increasing sampling and including brand category were suggested but these actions will cost money money to do so. With all the information and capital supply we have SVM model is the best option to replace the manual laboratory analysis by predicting CatSO3 levels in the whitebase production and it could be used by Procter&Gamble Company, Pomezai plant.

References

[1] W. C. Dampier, “The Encyclopaedia Britannica,” *Science*, vol. 8, p. 402, 1911.

[2] J. Li, X. Tong, L. Zhu, and H. Zhang, “A machine learning method for drug combination prediction,” *Frontiers in Genetics*, vol. 11, 8 2020. [Online]. Available: https://doi.org/10.3389/fgene.2020.01000

[3] Z. Wang, Z. Sun, H. Yin, H. Wei, Z. Peng, Y. X. Pang, G. Jia, H. Zhao, C. H. Pang, and Z. Yin, “The role of machine learning in carbon neutrality: catalyst property prediction, design, and synthesis for carbon dioxide reduction,” *eScience*, vol. 3, no. 4, p. 100136, 8 2023. [Online]. Available: https://doi.org/10.1016/j.esci.2023.100136

[4] I. P. Corporation, “An easy guide to understanding how surfactants work,” 2022. [Online]. Available: https://www.ipcol.com/blog/an-easy-guide-to-understanding-surfactants/

[5] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. Springer Science Business Media, 5 2013.

[6] M. Mitchell, “Selecting the correct predictive modeling technique,” 9 2022. [Online]. Available: https://towardsdatascience.com/selecting-the-correct-predictive-modeling-technique-ba459c370d59

[7] H. Zou and T. Hastie, “Regularization and Variable Selection Via the Elastic Net,” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 67, no. 2, pp. 301–320, 3 2005. [Online]. Available: https://doi.org/10.1111/j.1467-9868.2005.00503.x

[8] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the royal statistical society series b-methodological*, vol. 58, no. 1, pp. 267–288, 1 1996. [Online]. Available: https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[9] “LaSso Regularization of Generalized Linear Models - MATLAB Simulink.” [Online]. Available: https://www.mathworks.com/help/stats/lasso-regularization-of-generalized-linear-models.html

[10] “Understanding Support Vector Machine Regression - MATLAB Simulink.” [Online]. Available: https://www.mathworks.com/help/stats/understanding-support-vector-machine-regression.html

[11] L. Breiman, *Classification and regression trees*. Chapman Hall/CRC, 8 2017.

[12] “Neural Networks: What are they and why do they matter?” [Online]. Available: https://www.sas.com/en_sa/insights/analytics/neural-networks.html#:~:text=Neural%20networks%20are%20computing%20systems,time%20%E2%80%93%20continuously%20learn%20and%20improve.

[13] “JMP help.” [Online]. Available: https://www.jmp.com/support/help/en/17.2/index.shtml#page/jmp/jmp-documentation-library.shtml

[14] H. Shimodaira and T. U. of Edinburg, “Single Layer Neural Networks,” 3 2015. [Online]. Available: https://www.inf.ed.ac.uk/teaching/courses/inf2b/learnnotes/inf2b-learn-note11-2up.pdf

[15] J. R. Koza, D. Andre, F. H. Bennett, III, and M. A. Keane, *Genetic Programming III*. Morgan Kaufmann, 1 1999.

[16] K. B. Prakash, *Data Science Handbook*. John Wiley Sons, 9 2022.

Acknowledgments

Special thanks to:

- my family: Serpil Albut and Selçuk Albut for supporting me on this journey and also supporting me everytime.
- my supervisor Professor: Alberto Roverato for guiding me patiently and teaching me the statistics.
- my school: University of Padova for giving me the necessary knowledge to complete this journey.
- my supervisor in P&G: Francesco Porpora for making this research possible.
- my team and my company in P&G: For helping me in every step with their process knowledge and friendship.