NonTechnical Report

```
getwd()
```

```
## [1] "C:/Users/user/Documents/Dec13NonTechReportCDrive/Dec13NonTechRptCMydoc"
```

INTRODUCTION

PROJECT 2

This is role playing. I am your new boss. I am in charge of production at ABC Beverage and you are a team of data scientists reporting to me. My leadership has told me that new regulations are requiring us to understand our manufacturing process, the predictive factors and be able to report to them our predictive model of PH.

Please use the historical data set I am providing. Build and report the factors in BOTH a technical and non-technical report. I like to use Word and Excel. Please provide your non-technical report in a business friendly readable document and your predictions in an Excel readable format. The technical report should show clearly the models you tested and how you selected your final approach.

Please submit both Rpubs links and .rmd files or other readable formats for technical and non-technical reports. Also submit the excel file showing the prediction of your models for pH

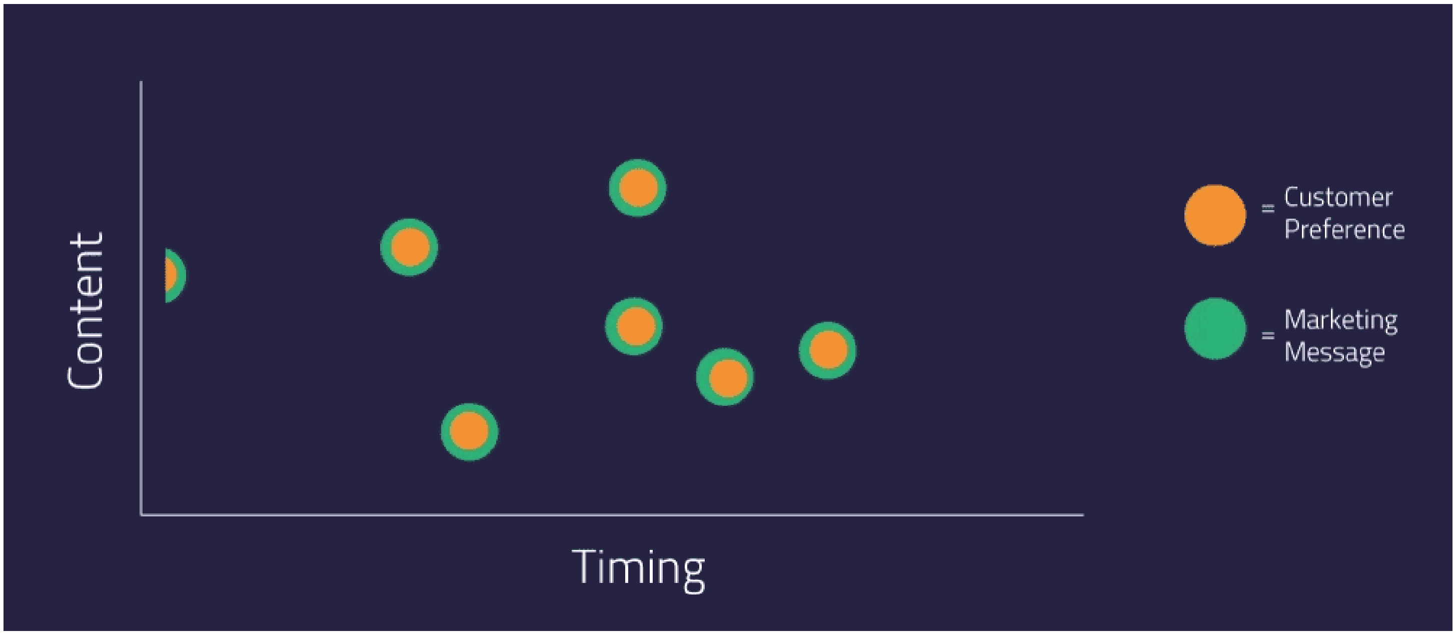
BACKGROUND

What is the top pain point for business executives? The clear answer among top business leaders: volatility. Too many factors — from regulatory changes, weather fluctuations to posts by social media influencers — impact supplier chain disruption, an even to buyers frequently change their minds.

Worse still, things reshaping business stability happen quite unexpectedly, such as stay from home order caused by covid, and in this assignments case, by new regulatory. But the uncertainty goes much more beyond these factors. On top of them, one side of change such as consumer taste change, or their newly gained social consciousness also impact other aspects of the whole business chain, which further trigger more regulatory changes and supply chain shift.

BUSINESS-A DYNAMIC REALITY

MARKET NEED /REGULATORY UNCERTAINTY



Figure

There is no magic wand to predict scenarios like the “COVID e□ect”, or what’s coming up in legislatures mind. We are living in such a quick and the unpredictable social an environmental society that our past knowledge is not enough if we rely on them alone. But there are technologies to improve the accuracy of demand forecasting. Honestly, it will never be 100 percent precise, yet it can be precise enough to help you achieve your business goals.

Demand planning (or Demand forecasting) , according to the Institute of Business Forecasting and Planning (IBF) applies “forecasts and experience to estimate demand for various items at various points in the supply chain.”

Demand planning serves as the starting point for many other activities, such as warehousing, shipping, price forecasting, and, especially, supply planning that aims at fullfilling the demand and requires data on the anticipated needs of customers. efficiency across the entire supply chain.

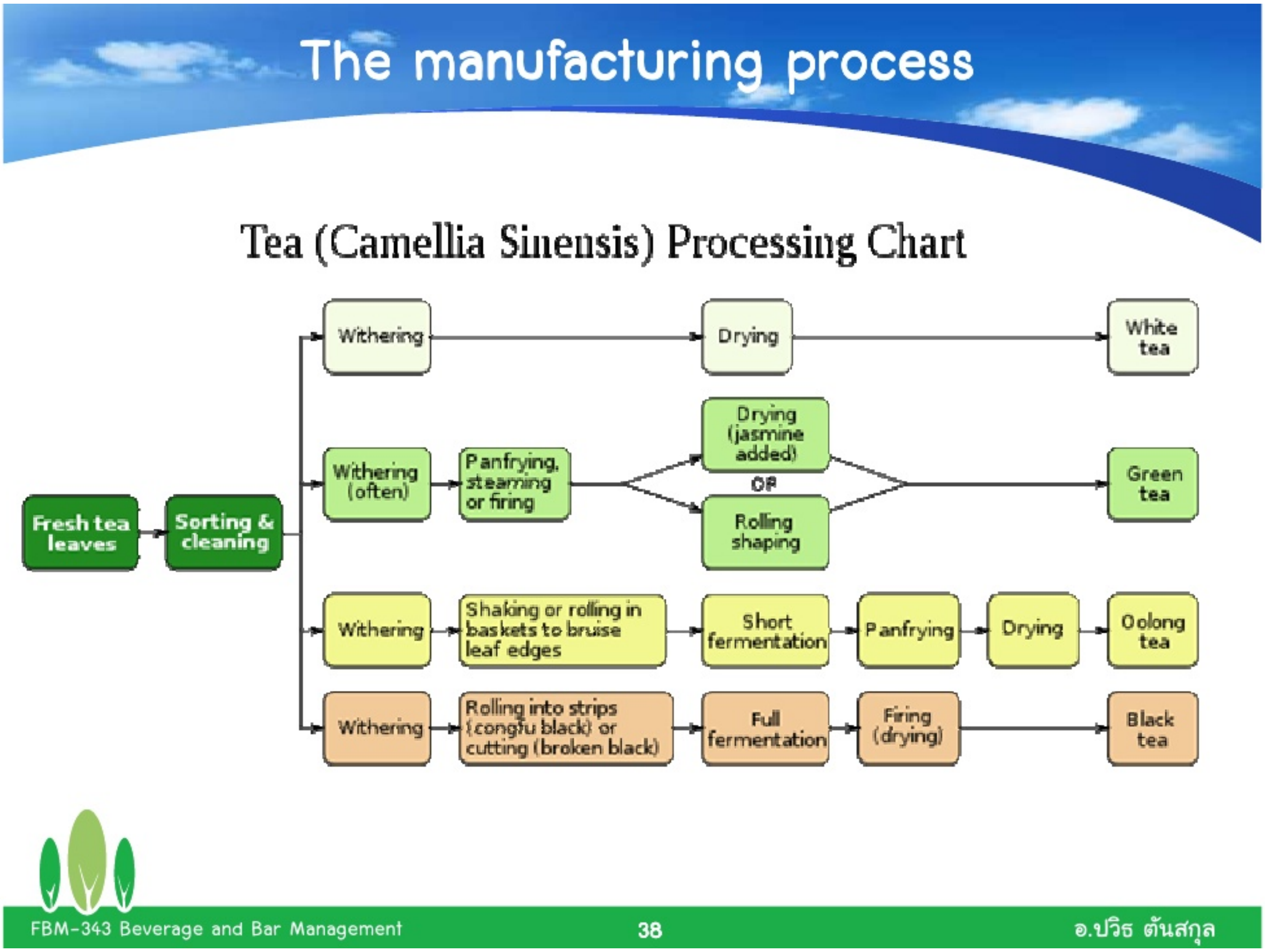
BEVERAGE PRODUCTION PROCESS

While regular consumers might have taken advange of the long standing beverage labels such as Coco Cola brand. But today’s market is such a dynamic market that consumers not only take into consideration the taste, the brand reputation, but also many many other factors such as nutritional benefit, social and environmental impact such as pollution of the production process to the global warming,waste of bottle to the sea pollution, so on so forth.

These are the main steps in the creation process: Formulation , Mixing , Flavoring , Testing , Packaging , Labeling .

Just for Nutritional Beverages alone, there are below common types for ingredients: Infused waters ; Sparkling juices ; Carbonated drinks ; Fruit juices ; Tea mixtures ; Ready to Drink Protein Drinks

Below graph shows tea manufacture process as an example:



The manufacturing process

Tea (Camellia Sinensis) Processing Chart

Fresh tea leaves → Sorting & cleaning

Withering → Drying → White tea

Withering (often) → Panfrying, steaming or firing → Drying (jasmine added) OR Rolling shaping → Green tea

Withering → Shaking or rolling in baskets to bruise leaf edges → Short fermentation → Panfrying → Drying → Oolong tea

Withering → Rolling into strips (congfu black) or cutting (broken black) → Full fermentation → Firing (drying) → Black tea

FBM-343 Beverage and Bar Management 38 อ.ปวิธ ตันสกุล

Figure

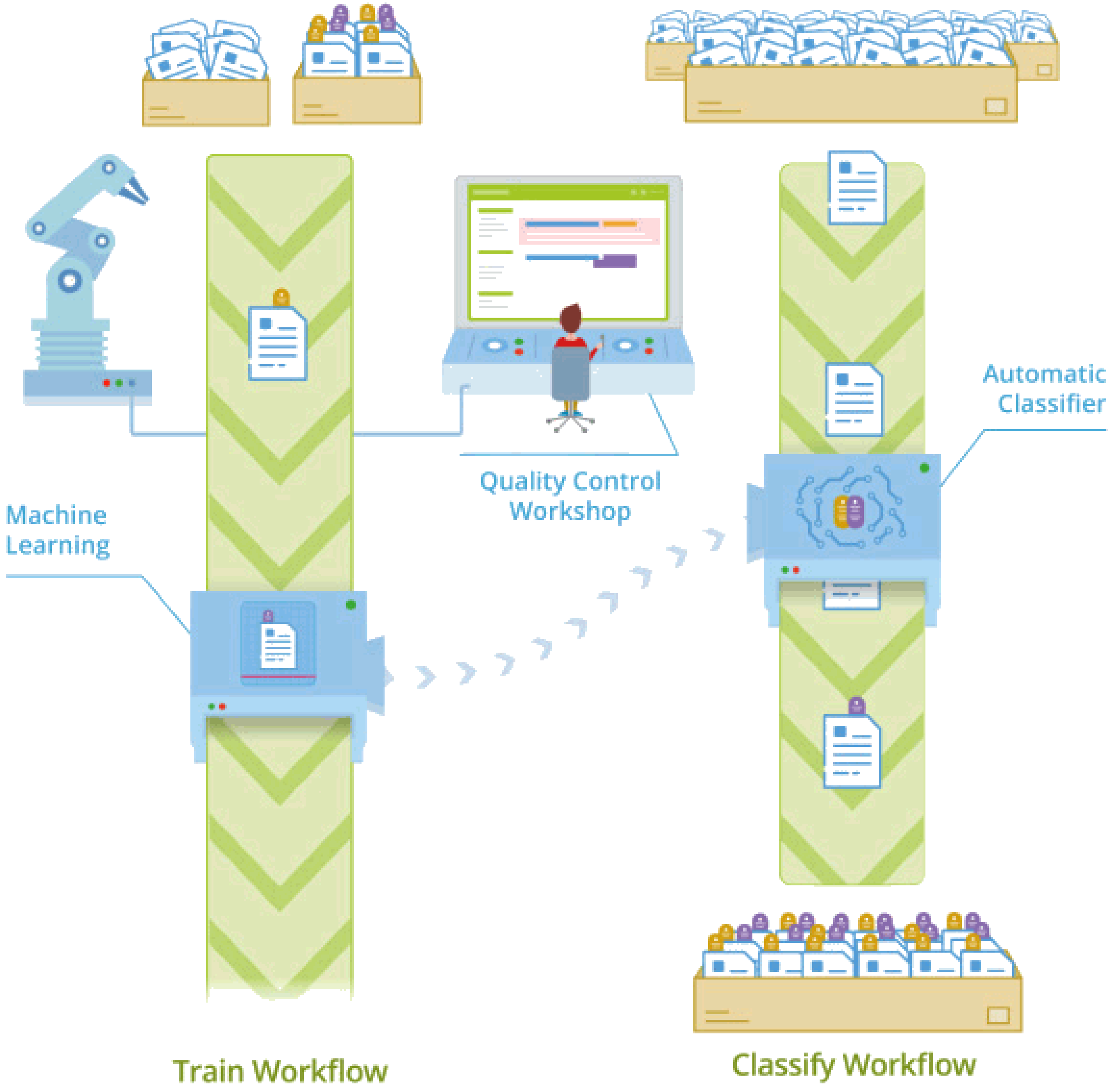
PURPOSE– MINIMIZE REPETIVE WORK

MACHINE LEARNING APPLICATION NECESSITY

So while it is a great idea for one or more beverages to sell, the reality of putting together all the pieces to bring ideas to reality can be daunting. The many steps required to bring a beverage formulation to market consists of ideation, R&D, formulation, flavoring, and manufacturing. What we do is remove all of the headache in managing these steps to deliver consumers beverage that they like, but also feasible for company to make a profit, by focusing on designing and marketing them, rather than other nuance factors.

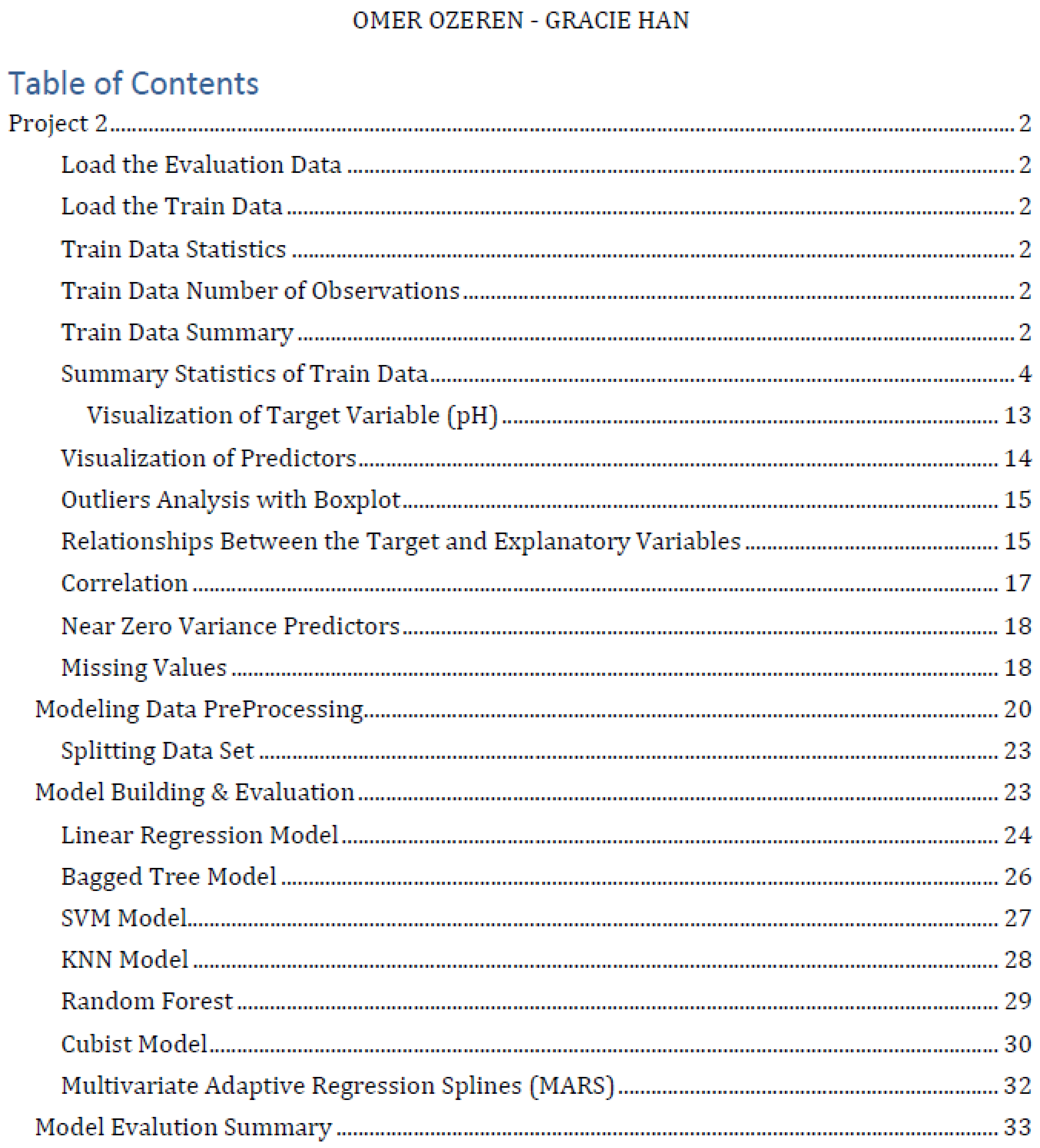
For this project, we intend to use the machine learning technology to for a specific research question: what are the factors that influence the pH outcome in ABC company’s manufacturing process? Our goal is to utilize historical data, put them into machine learning process, find the patterns that may influence the outcome, and then modify them via machine learning algorithm, and then to predict future PH outcome.

By doing so, just like the below graph shows, we want to automate what machine can do for us in decision-making process, and let the treasured human mind focus on the crucial task, such as designing and sale. In other words, many of the QC intermediate steps, which are now analyzed by individual human brains, can be assisted by the mastermind of computers. Current stage of modern machine learning and forecasting models have made such task a reality.



## METHODS

Detailed technical process, starting with how the data are composed, how we process the data, how we apply the machine learning models into them are detailed in this technical report. I am not going to elaborate in detail what I have already put into the technical report. Rather, we are highlighting the part that require the thinking process and stimulating further discussion for future research in manufacture process.



OMER OZEREN - GRACIE HAN

Table of Contents

Project 2 2

Load the Evaluation Data 2

Load the Train Data 2

Train Data Statistics 2

Train Data Number of Observations 2

Train Data Summary 2

Summary Statistics of Train Data 4

Visualization of Target Variable (pH) 13

Visualization of Predictors 14

Outliers Analysis with Boxplot 15

Relationships Between the Target and Explanatory Variables 15

Correlation 17

Near Zero Variance Predictors 18

Missing Values 18

Modeling Data PreProcessing 20

Splitting Data Set 23

Model Building & Evaluation 23

Linear Regression Model 24

Bagged Tree Model 26

SVM Model 27

KNN Model 28

Random Forest 29

Cubist Model 30

Multivariate Adaptive Regression Splines (MARS) 32

Model Evalution Summary 33

Figure

OUR DATA

we call them predictors. Among them, Brand Code has 4 categories, while the rest of the 32 predictors have very detailed information, in continuous variable and numerical variable prospect 2 describe the detail the manufacture process.

```
[1] “Brand.Code” “Carb.Volume” “Fill.Ounces” “PC.Volume” “Carb.Pressure” “Carb.Temp”
[7] “PSC” “PSC.Fill” “PSC.CO2” “Mnf.Flow” “Carb.Pressure1” “Fill.Pressure”
[13] “Hyd.Pressure1” “Hyd.Pressure2” “Hyd.Pressure3” “Hyd.Pressure4” “Filler.Level” “Filler.Speed”
```

```
[19] "Temperature" "Usage.cont" "Carb.Flow" "Density" "MFR" "Balling"
[25] "Pressure.Vacuum" "PH" "Oxygen.Filler" "Bowl.Setpoint" "Pressure.Setpoint" "Air.Pressurer"
[31] "Alch.Rel" "Carb.Rel" "Balling.Lvl"
```

For the sample size limitation, we did not look at the brand individually. Rather we pulled all the brands together and looked as them as a wholesome analysis.

Our goal is to predict the PH value, we call that outcome, utilizing most or all of these 33 variables.

QUALITY OF DATA / PREPROCESSING OF DATA

We also looked at the outcome, pH value, and found that it is normally distributed, which means that it can be utilized as is without further manipulation for machine learning process. This is very important, because if the outcome is not trustworthy in data quality, no machine can overcome the difficulty of bad data.

In order for us to make meaningful extrapolation of the historical information, we need the data to be in reasonably good format, so that the forecast are not far away from the reality. So we first looked at the quality of the data.

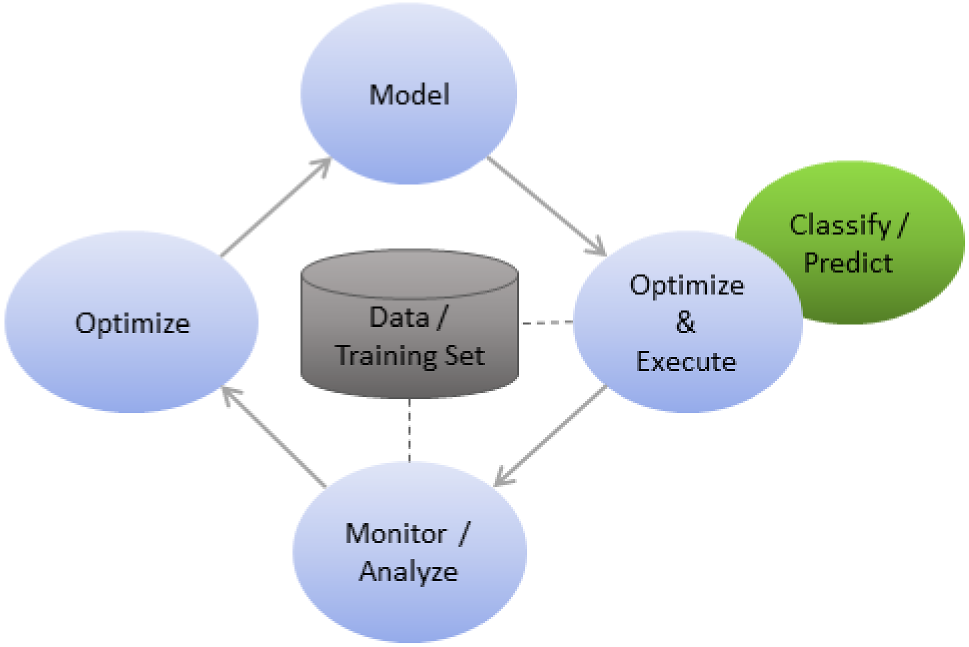
First, sample size, we were given a total of 2571 observations of historical data, which is sufficient for machine learning purpose. Next, we looked at the missing data amount within these 2571 observations. We found that there are about 8% of the observations have some missing data, which is reasonable for most of the datasets. And more importantly, dismissing values are at random, which is important that they will not influence significantly the outcome.

For the missing data, we further investigated where do they occur and to which predictors do they occur. We also imputed most of them using machine learning algorithm. The purpose of imputation is to preserve the most of the observations we have had, and utilizing their information to the maximum. If we did not impute, then the observations that has any missing values will likely be thrown out, therefore 1 predictors missing will affect the rest of 30 predictors prediction power., however, certain variable, Hyd.Pressure1, we chose not to impute it's missing, due to the pattern of it's missing mostly at 0, which violates the rule of thumb for imputation.

After the imputation, we have preserved most of the data for the machine learning process.

MACHINE LEARNING PROCESS

This below figure is a general illustration that summarizes the rule of thumb for machine learning process. Regardless of which machine learning algorithm we chose, they all followed the below process.



Figure

First step, the data is preprocessed, including inspecting its quality of predictors and the quality of outcomes. Also include inspecting the patterns of predictions to outcome, many times individually, meaning that the individual predictor association to the outcome are looked at 1 by 1, and the individual predictors association among themselves are also looked at one by one. Only the meaningful information, oftentimes decided by human brain, are utilized.

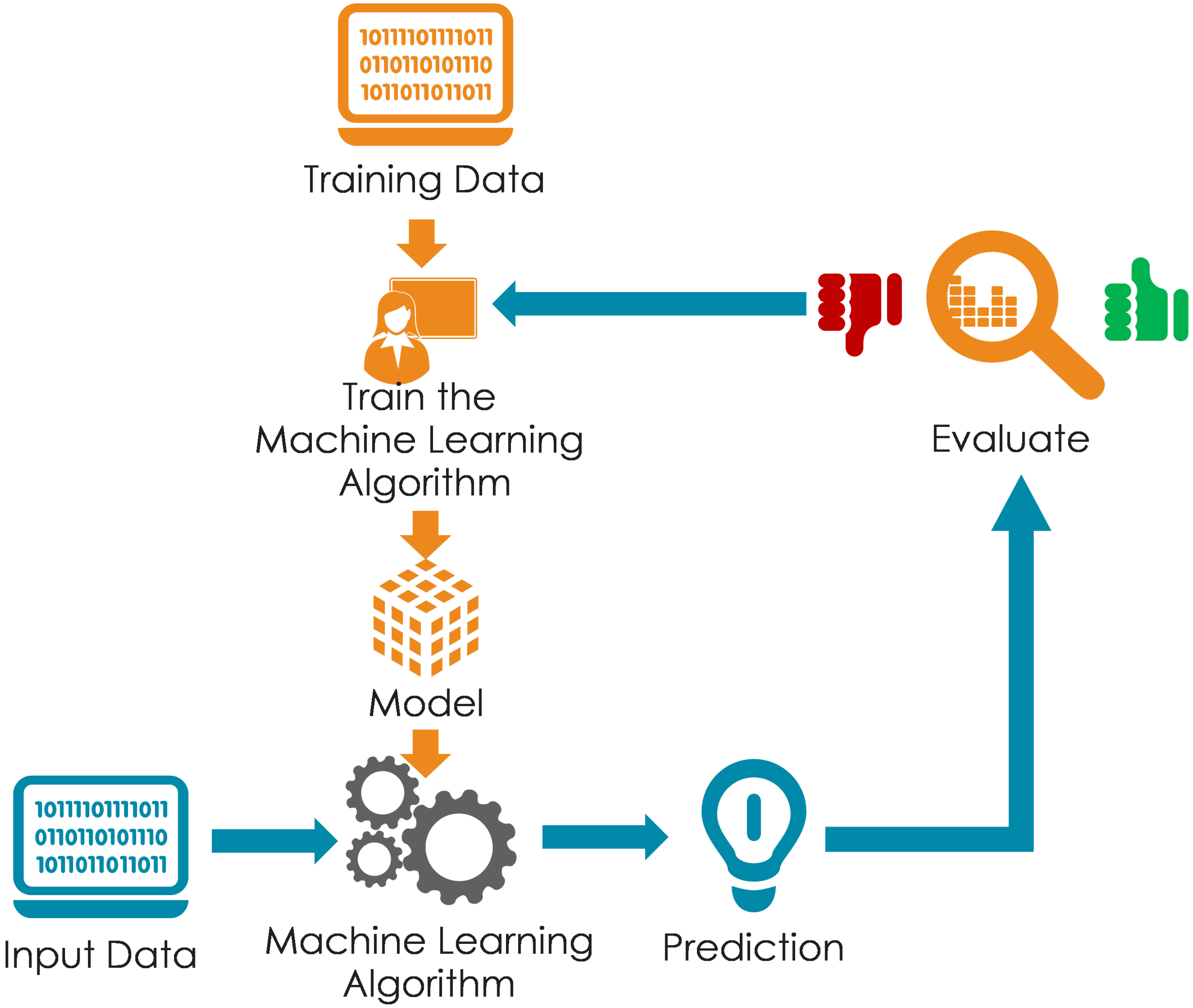
Second step, the bad quality data, including missing data, outliers, are preprocessed, if possible by imputation and other mechanisms, so that maximum information yet still relatively accurate information are preserved. Then the fun process of machine learning starts there.

Third step, to split the data. For any machine learning, the data is split into 70% , by choice, into training data set. The rest of the data is put into test data set., which are a random process, meaning the data are shuffled in terms of their position, to introduce maximum randomization and to make the algorithm more accurate.

4th step, certain machine learning algorithms are chosen to apply to this data set. there are many default choices within each algorithm that computer generates, but also on top of that, many parameters can be audited and changed, to fit specific data. These parameter change are up thinking process of human brain, to supplement the machine learning algorithm automating process.

5th step, different machine learning algorithms of choice produce their output after its training and testing process. The predictors for overall machine learning algorithm, on how well they fit this data can be exported and compared together.

six step, by comparing the evaluators across the machine learning algorithm for this particular data, the optimum model for this particular training is chosen. Then the final prediction on two individual observations of the testing data are exported. All of the 33 predictors can be predicted, in terms of their individual variable number on all 30 three of them. This will serve as an intuitive and meaningful information for legislatures to inspect and make recommendations for future manufacture improvement purpose. And that is also our goal and purpose of this research project.



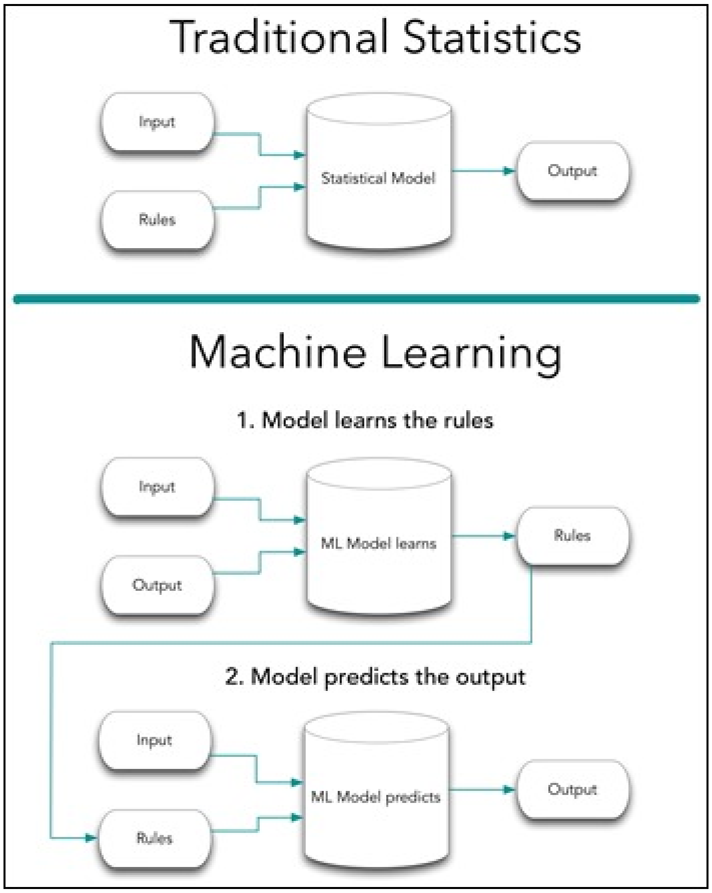
Figure

MACHINE LEARNING VS TRADITIONAL STATITCS

You may wonder, why machine learning? In the past, I rely on statistics to extract and extrapolate business information, which serves the purpose very well for my business knowledge. Many of your machine learning algorithms cannot be explained well, and how would I trust a computer better than the truth (statistics)? As a business leader, I do not like any random machine who cannot explain themselves well to tell me what to do, I like truth and nothing but truth.

Well, I can certainly understand where you come from, but I would like to give you this introduction on statistics and machine learning. Please be aware that they are not contradicting each other, rather they are complementing each other.

Below figures illustrate the crucial process in statistics and the machine learning.



Figure

As illustrated from that figure, there are many rules that statistics and machine learning apply as common rule. As a matter of fact, all of the machine learning models start from established statistical model. And as machine learning has progressed in terms of its sophistication, it still can be traced to its statistical path. Also, all of the validation mechanism are derived from statistics. In other words, we can safely say that statistical model and the traditional well established statistical input and output including its rules, are well integrated into most of the machine learning algorithm.

What machine learning shines is at below:

1, Machine learning model learns the rules, which on most cases are statistics based rules:

2, Machine learning model predicts the output. One footnote here, machine learning model, do not just randomly predicts the outcome. Rather, it follows the statistical pattern that the first step has recommended.

In conclusion, machine learning integrates the statistical path, rather than deviate itself from statistics. So, machine learning is based on facts, rather than randomness. Of course, there are some caveats, which is beyond the topic of our report today.

CONCLUSION

COMPARISON- RULE BASED ML/TRE BASED ML / TRADITIONAL ML

We have applied below machine learning (selective) to our data:

Linear Regression Model

Bagged Tree Model

SVM Model

KNN Model

Random Forest

Cubist Model

Multivariate Adaptive Regression Splines (MARS)

First, we run linear regression model. This is our basic banhmark model.Because linear regression is a traditional model, and our data contains mostly numerical continuous variable, and our outcome pH is also continuous variable. Therefore we first chose linear model as the basic machine learning technique to predict the beverage’s PH outcome.

The next few models are all assuming non-linear fashion, which are more popular machine learning algorithms and also more truthful to this data prediction. We chosed a few treebased modeling.

The ensemble techniques of the nonlinear models have a few advantages. By packing or bagging the variables into trees,the variance of a prediction through these ensemble process are reduced, which fit even the unstable predictions with less stringent assumption than linear model.

Lastly, we applied the rule based model cubist.

For this particular research project, we found that cubist model has the best performance overall for this research project. Therefore for our final output, we produced the individual predictions for all 33 predictors, in Excel format for the legislations to consider.

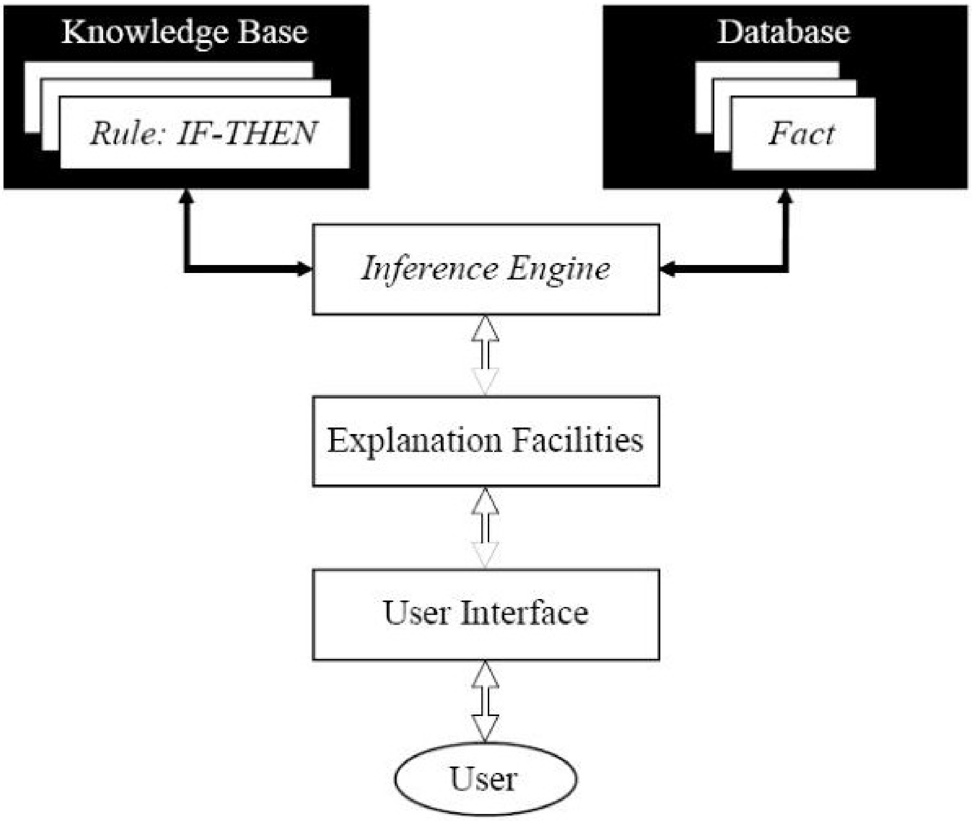
DISCUSSION

RULE BASED MACHINE LEARNING MODELS

For this particular research project, we found that cubist model has the best performance overall for this research project. Therefore for our final output, we produced the individual predictions for all 33 predictors, in Excel format for the legislations to consider.

Because cubist machine learning model clearly outperforms the traditional linear regression machine learning model, also outperforms all series of tree based models, we expect to use this machine learning model for many more future analysis for our data. That’s why we are illustrating in detail the concept and the future directions for this research.

Basic structure of a rule-based expert system



Knowledge Base

Rule: IF-THEN

Database

Fact

Inference Engine

Explanation Facilities

User Interface

User

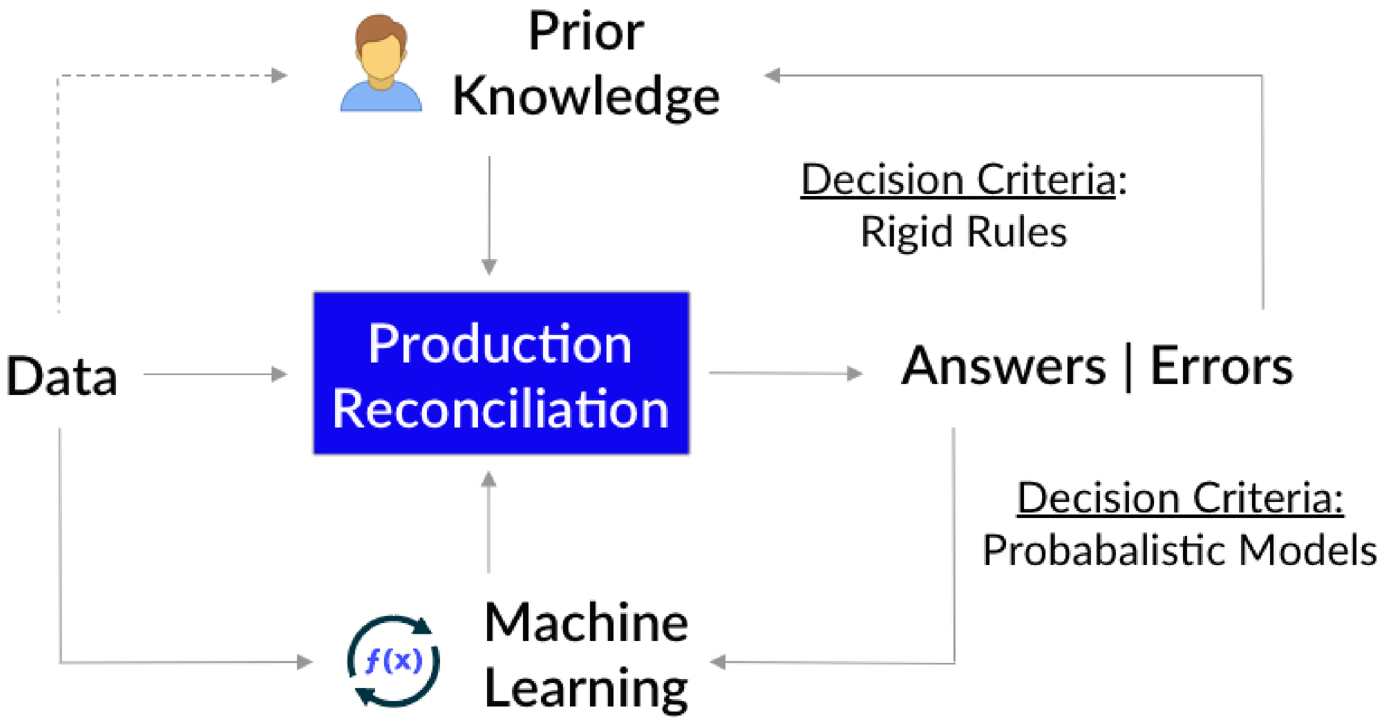
Figure

Cubist is a rule based machine learning model. So this figure illustrates what a rule-based model is doing and why it outperforms other models in our setting. In short, for data base that has many predictors come up versus straight uncomplicated situations like 2 predictors, the traditional linear model can handle only small sets of situations. Then, the tree-based models rely on too many randomization of the machine learning, which makes its prediction power small.

Rule based models, such as cubist, shines in modern big data.

Furthermore, the so called the rules our manipulable manipulate can be manipulated further. there are parameters within this package that can be further modified, which are so called committees, an neighbors, I am not going to dig too much into detail technically.

NEED OF “HUMAN RULES”


Prior Knowledge
Decision Criteria: Rigid Rules
Data
Production Reconciliation
Answers | Errors
Decision Criteria: Probabalistic Models
f(x)
Machine Learning

Figure

so now, the big question is are we totally succumb to machine? Machine learning is powerful great quick precise and everything. do we need human brain injection?

the answer is loud and clear: yes we do! Now we are only at the tip of iceberg of integrating machine learning and human knowledge. The so called the training of the machine relies on historical data, and it’s algorithm are being optimized and developed on a daily basis. We have to understand that not all algorithms apply to all specific field, such as production of carbon hydrate drink. The subject knowledge, many times intuitive to human, but foreign to machine are crucial in successful product development.

The so called rule based model, while it applies to machine, I will not be surprised that in the near future, it can be manually interrupted by human rules. This below illustration is a flow chart of human rules. I would like to conclude by saying that machine learning is progress in, so is our human knowledge. For future projects, we recommend that the machine learning rules and human rules are combined together.