Winery Fermentation Analytics

Asad Mahmood, Irene Ramirez, Trang Vu, Laura Bathurst, Zhonghan Deng, Jared Halterman, Saba Sajjadi, and Moxshil Shah

M.S., University of the Pacific, 2021

Dr. Emma Hayes
Advisor

A Capstone Project Presented to
The Graduate Program in Partial Fulfillment of the Requirements
For the Degree of Masters in Data Science

TABLE OF CONTENTS

INTRODUCTION

During the Prohibition, two grape growers, Ernest Gallo and Julio Gallo, sold their California grown grapes to Eastern states, where home winemaking was permitted. When Prohibition was repealed in 1933, the two grape growers decided to skip the middleman and make the wine themselves (Jamieson, 2007). That year, E. & J. Gallo Winery was founded and is now renowned as being the largest exporter of California wine (McGowan, 2017) and largest family-owned winery in the world (IBM Research Blog, 2017).

Gallo has long been at the forefront of the data revolution in the wine industry, beginning with sophisticated analytics in sales and distribution in the late 1990s and extending into areas such as pricing and customer behavior as they were able to accumulate more data (Henschen, 2012). By 2011, Gallo was using deep data on consumer taste preferences and grape varieties to create a new and highly successful wine brand, Apothic (Henschen, 2012). In 2012, Gallo partnered with IBM to create a machine-learning based, IoT technology that integrated data from satellite imagery with vineyard sensors to automatically predict and regulate the irrigation of individual vines (IBM Research Blog, 2017). By 2017, Gallo had used data to transform harvest (McGowan, 2017). This semester, Spring 2021, Gallo sponsored a Capstone project that included our team, University of the Pacific Data Science Master's students to continue to expand their digital transformation applying predictive modeling to wine fermentation.

PROBLEM STATEMENT

The goal of the Capstone project is to determine when fermentation is complete and when wine is ready to be bottled. The time to completion would depend on the type of wine and location of the tank. The color of wine determines the required fermentation length, therefore different models must be implemented to increase the accuracy of fermentation complete time. The benefit of this knowledge would allow Gallo to maximize their production and tank turn over time and minimize the labor of physical wine testing. Gallo shared their data collected on fermentation in hopes of acquiring an accurate prediction.

PROJECT TIMELINE

Our project began on January 15 with the first virtual meeting. We met with Joshua and John (Gallo contacts) and discussed Gallo itself and the goal of the project. Joshua also introduced the general wine making process. We also received basic information regarding what we will need to access Gallo's database such as Oracle SQL Developer, a VPN, etc. After our first meeting, we ran into technical issues restricting our team from accessing the Oracle database due to VPN issues. On February 4th, we had our first technical troubleshooting meeting and began working with Gallo to resolve these problems. Two weeks later, on February 18th, the team was able to gain access to the Oracle Database. We were able to inspect the database, however, we were unable to export data due to very slow processing speed.

As a result, Joshua began working on getting us access to the HANA database. On March 4th, our team was able to gain access to one HANA account, which gave us the ability to extract data at a much faster rate. However, before we could pull the relevant data required to build our model, we still needed more clarification on the relationships of different data tables in conjunction with the fermentation process. On April 1st, we were able to get clarification from Jessica and on April 10th our team was able to extract the necessary data. We then started to merge and clean the data for model use. On April 16th, the first models were developed and worked on. The insights and predictive modeling was done and the presentation planning began.

Finally we delivered our final presentation on April 26th. Although the timeline unveils our substantial struggle with data access and extraction, we were able to produce a passable model that revealed we are on the right track in selecting relevant features to predict the fermentation process. Figure A provides a quick overview of the Capstone project timeline.

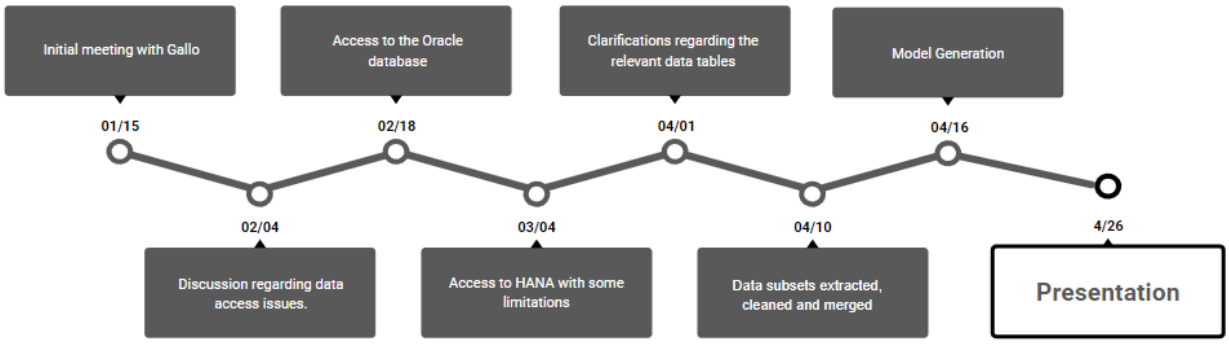


Figure A

PARTICIPANT ROLES AND CONTRIBUTION

The team tasked to collaborate with E. & J. Gallo winery consists of eight data science graduate students from University of the Pacific. All members of our dedicated team participated in team meetings with and without Gallo, collaborated using our team Slack workspace and Google Drive folder, and assisted in the creation of team deliverables, including progress reports, the project presentation, and the final project report. All team members also explored the complex data stored in Oracle and SAP HANA to achieve an understanding of the information available from Gallo and its limitations. Additional individual contributions are as follows:

1. Asad Mahmood:
    a. **Project Lead/Manager:** I created and assigned tasks through a document which was used by team members, such as merging separate data extracts, getting local temperatures, extracting wine ingredients and composition data, etc.
    b. **Data Scientist:**
        i. **Documentation and visualizations:** Drafted progress reports, presentations, final reports, and created all the visuals used for graph database and predictive analysis explanation throughout the duration of this project.
        ii. **Query Building and Data Extraction:** Initially, we got access to the SQL developer, I built queries and searched for relevant tables and columns in them. But SQL developer didn't allow us to export data, so we were given access to SAP HANA Views where I repeated the same process and was able to extract data regarding different measurements taken during wine sampling such as NOPA, Ammonia, Glucose/Fructose, etc. as csv files. These csv files were used by Zhonghan Deng to create the initial dataset and later on by Trang Vu to create the final dataset.
        iii. **Predictive Analysis and Code Development:** Initially, I devised the modeling strategy around time series forecasting using both Tensorflow and FbProphet (An open source time series forecasting framework developed by Facebook). But after getting access to HANA views and successful data extraction, I went with regression analysis after taking into account we had just 296 records which were taken inconsistently so time series forecasting was out of the question. I read several research papers on wine making relating to the fermentation process and narrowed down what other features were necessary for predicting fermentation other than ones already in the initial dataset. I shared my findings with the team

and Trang Vu worked on it and created the final dataset. I trained both the initial and final dataset on thirty plus regression algorithms to select the best performing one, which was ExtraTreesClassifier, with the final dataset showing better results overall compared to the models trained on the initial one.

c. **Troubleshooting:** I helped fellow peers troubleshoot data access and installation issues by providing detailed tutorials. Such as installation of TNSPing, SQLPlus installation and requirements, setting up HANA Views using DBeaver/Eclipse, etc.

2. Irene Ramirez:
    a. **Project Management:** Communicated with both E&J Gallo and Dr. Hayes, middle-man who prepared and edited emails and reports. Delivered final reports in Canvas and kept all parties up to date on project's progress and meetings.
    b. **Data Scientist:**
        i. **Query Building:** Initially explored tables in Oracle and then assisted with the query building and development in Hana, as well as prepared and edited new queries for data exploration. Used the final queries built and created them in python for a seamless process.
        ii. **Data Extraction:** Completed data cleaning, data preparation, data manipulation for final model using HANA and python.
        iii. **Code Development:** Developed and compiled final code to provide a comprehensive package to the client, including files and code.
3. Trang Vu:
    a. **Data Scientist/Engineer:**
        i. **Coordinating**: Communicated back and forth with E&J Gallo experts to identify issues and gain accessibility of data. Communicated with Gallo experts about the usefulness of certain tables for ingredient exploration, gaining clarity in terms of query composition and details of relevant data that is highly correlated to the fermentation process by following up with them about winemaker questions discussed with Asad. Assisted the group in drafting reports and final presentations. Kept Dr. Hayes and Gallo posted with the project's progress and followed up on scheduling meetings.
        ii. **Accessibility troubleshooting**: Assisted other Mac users - teammates to solve the connectivity problems to Gallo's database through SQL developer/ Hana by fixing the VPN/ security issues with Mac OS by installing anti-malware on systems in order to be compatible with Gallo security policy. Facing accessibility issues to the

ingredient tables in Hana, I also communicated with Gallo IT specialists and DBA to identify resolution and troubleshoot access.

iii. **Query building and data extraction/preparation**: I developed multiple joins queries to extract ingredient material compositions and selected columns of interest through Hana. I then merged the exported ingredient dataset to the initial dataset using python. I developed and compiled code to filter final combined dataset by only rows that have the latest completed work orders of ingredient additions that happened before sampling for specific tanks and lots. Thus, I performed data manipulation, data cleaning and one-hot encoding on categorical columns to be ready for model implementation.

iv. **Data exploration and analysis**: Along with code development, I explored and analyzed ingredient dataset to find key ingredients for the fermentation process. I developed visualizations to explore the distribution and consistency level of how sample data collected versus different measurements in the final dataset. Also, tested/manipulated model implementation code and assisted with final combined code testing and compiling to be ready to deliver to the client.

4. Laura Bathurst:
    a. **Data Scientist**:
        i. **Query creation and refinement:** I experimented with many alternative query constructions in hopes of either finding or integrating tank temperature data into our dataset or else extracting data appropriate for creating an event sequence model.
        ii. **Fermentation domain knowledge:** I communicated with Gallo regarding the importance of access to domain expertise about the fermentation process, in response, they shared useful one-page diagrams of white, red, and sparkling fermentation processes. I also shared relevant online sources for understanding wine fermentation with the team, including a 2002 Gallo manual I located online targeted at explaining Gallo processes to sales people that—despite being almost 20 years old—contained useful information to help decipher Gallo's SQL tables. Despite being dated, it gave possible hints as to the complex and distinct processes for different kinds of wines that Gallo may be using, as well.
        iii. **Non-Gallo fermentation data:** I searched for data to use in place of, or as a supplement to, the data we were unable to get from Gallo. When I was unable to locate data of this type, I explored possible methods for creating synthetic data

suitable for our purpose and completed preliminary steps toward creating such data before turning to other priorities due to time constraints and team priorities.

5. Zhonghan Deng:
    a. **Data Scientist / Engineer:**
        i. **Data exploration / manipulation**: I did also explore our sample dataset to see if time-series models can be used. (Created different data visualization) But after my exploration, we as a group thought the data is not proper for a time-series predictive model. After getting all csv files from database, I did data combination, data cleaning, data preparation, data manipulation (Such as: handle NAs, generate new columns, etc.)
        ii. **Query Building**: Worked in HANA and SQL server to explore useful data that can be used in our predictive model. Extracted data from HANA in csv format.
        iii. **Troubleshooting**: Worked with Gallo IT to figure out and solve VPN connection issues.
        iv. **Non-Gallo data exploration**: Because we do not have temperature data provided from Gallo and they state temp is an important feature. So I was trying to find a public API for history local temp and use that as a substitute.
6. Jared Halterman:
    a. **Data Scientist:**
        i. **Query Building:** Worked in Hana to develop a query to pull data with a dummy variable for wine color (red/white). Required analysis of the data columns to determine the correct rows needed for our project.
        ii. **Data Extraction:** Extracted large amounts of data to share with group members.
        iii. **Data Table Exploration:** Spent too much time parsing through data tables looking for relevant information relating to our project.
    b. **Troubleshooting:** General troubleshooting.
7. Saba Sajjadi:
    a. **Data Scientist:**
        i. **Troubleshooting:** Communicate with Gallo to get access to data and troubleshooting VPN access for team members. Helped mac team members to troubleshoot Gallo VPN connectivity issues by installing Anti-Malware software which was compatible with Gallo security policy.

ii. **Data Extraction:** Analyzed tables in HANA to find important ingredients for the fermentation process and extracted the useful tables along with SQL and Python code development.

iii. **Query Building**: Used python to join different dataset and prepare the final dataset to be used in our final model.

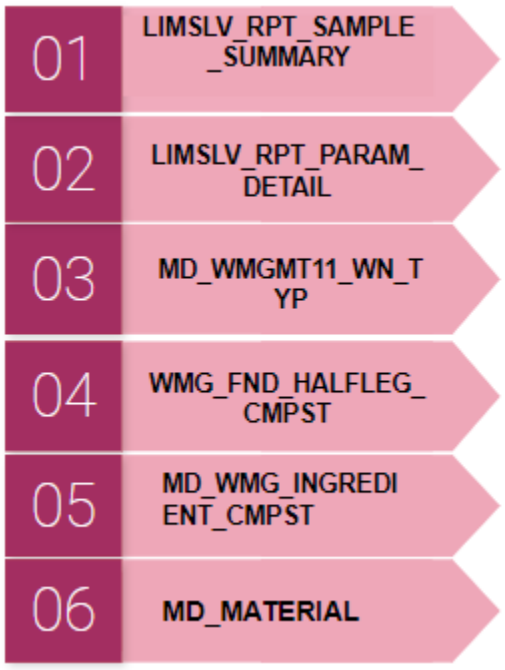
8. Moxshil Shah:
    a. **Access Troubleshooting -** Worked with Gallo to troubleshoot the VPN connection and SQL Access. I informed Gallo about the credentials for SQL, as mine was not created. Later, I still had trouble accessing the SQL Developer and SQL Plus. Gallo and other members of the team helped me to solve the issues.
    b. **Meetings:** I make a record of weekly meetings, and before the submission of a progress report, I share them with my members and discuss it on the Slack Group.
    c. **Submission:** As the Canvas Dashboard was not working properly to submit the file as group work. I immediately informed this issue to team members, and then wrote mail to our faculty advisor for a solution.

DATA OVERVIEW AND BRIEF EXPLORATION

Knowledge of and selection of relevant data was discovered through personal research (using research papers about the fermentation process and a Gallo sales manual from 2002), meetings with Gallo's experts, and data exploration. Our team identified available key components in the provided databases that impacts the fermentation process such as sulfur dioxide (SO2), lactic acid, and malic acid.

Gallo employees provided us with a basic understanding of how fermentation is done with a focus on their Livingstone facility. They walked us through the process of how the data sample was collected and logged into the LIMS database system, with one of the fermentation process states and different measurements. They also helped identify useful features from the measurements recorded at the time of the wine sample data collection. These measurements were later used in the creation of the initial dataset, including NOPA, internal tank temperatures, pH values and glucose/fructose. Due to the inconsistency of the internal tank temperature data, we pulled local temperature data in hopes of finding significance with each batch. Gallo also assisted us to specify the relationship among tables and how we could link the tables together to achieve the data that we need. Additionally, the query formula for ingredient composition is also clarified.

For exploration, we used Python to go over the material dataset and identify the top frequency of the ingredient groups added in the latest work orders. The following ingredients exhibited the highest frequencies as follows: additive, acid, enzyme, yeast and nutrient.



The data sources offered include SQL Developer and HANA. Due to the slowness and data extraction timeouts that occurred in SQL developer, we used the HANA Views. While a query and data export in SQL Developer would take a while to complete, HANA can complete the same query or data extraction in a shorter period of time.

Within HANA, we found key tables to pull relevant data. Figure B lists the names of the 6 different data tables we were able to pull from. The first table is the wine sample data for the Livingston facility, the second table is used to extract relevant features and labels for modeling. The third table is used to determine which LOTs represent which

wines. The fourth is used to extract composition ID information along with tank, work orders and LOTs while the fifth is to extract Ingredient ID. We need these latter two data tables since they are required for us to retrieve the data of interest in the sixth table MD_Material which contains ingredient additions that could be used as features for modeling.

Figure B

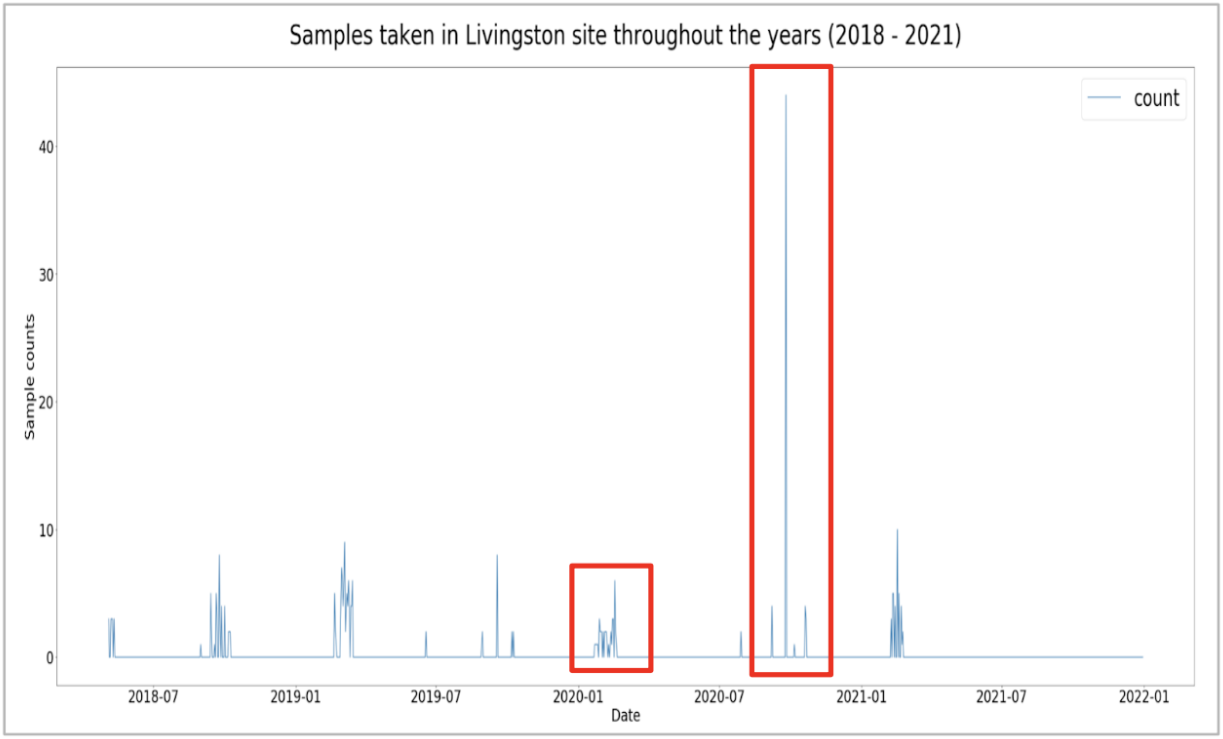


Figure C

After merging all the data collected from HANA, we got our final dataset with **36 columns and 296 rows**. The graph above, Figure C, shows the count of samples taken in Livingston throughout the year ranging from **2018 to 2021** based on the Sample Created Date column. The data samples were also taken inconsistently. Specifically, 2020 witnessed a high peak in terms of the amount of samples taken whereas the others period witnessed much smaller amounts of samples counts with non-constant amounts. In general, there is a contradiction between the high amount of features that we were able to extract and the low number of samples that we were able to collect.

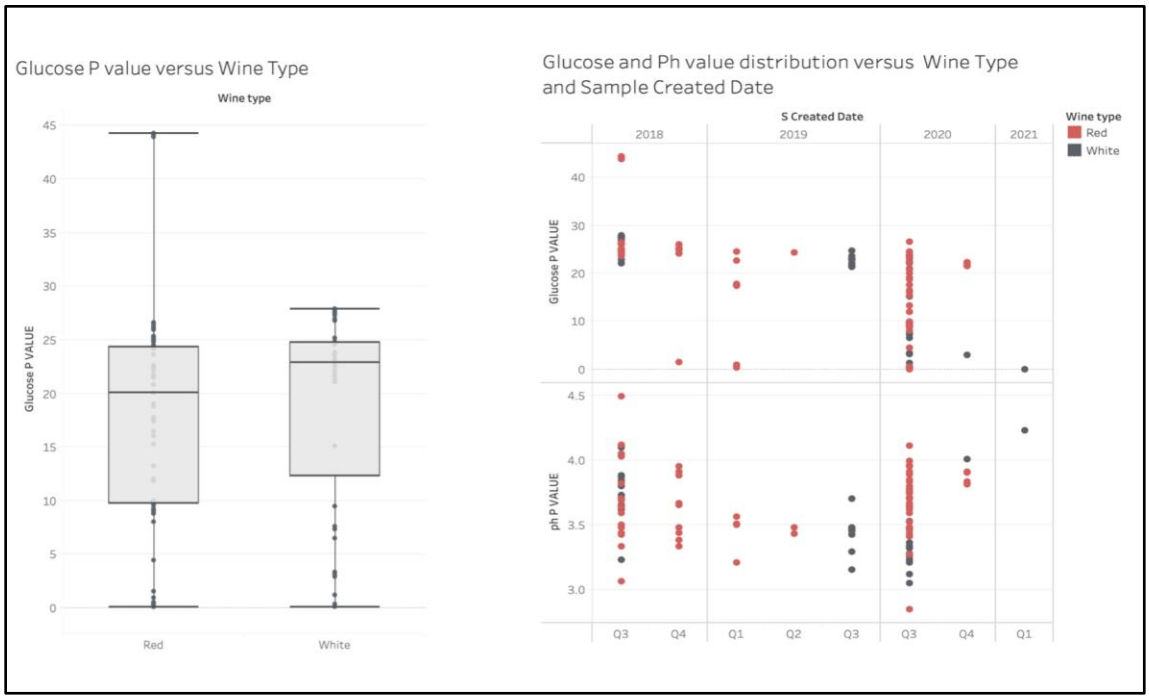


Figure C

Moving forward, we explored the data to understand the characteristics and distribution. The left side of Figure C illustrates the distribution of glucose p-value versus wine types. We could see that the distribution of glucose p-value in red wine is witnessed to be in a wider range than white ones. Meanwhile, the median of glucose p-value in white wine is higher than red ones. That is reasonable since white wine often contains more sugar than red wine and once the glucose value reaches certain low values then the fermentation is reached. Specifically, it often takes longer for red wine fermentation to be reached since facts have their own that red wine is often dry and less sweet than white. The plot on the right displays the distribution of two measurement pH values and glucose versus wine type throughout different quarters when samples were taken. It is noticeable that there are wider distributions of red wine than white wine data throughout the plot. In the last quarter of 2018 and the first two quarters of 2019, we only see the sample of red wines. In the third quarter of 2018 and 2020, there are more evenly and dense distributions of red and white. Especially, we could also see some quarters with only white wines as samples taken. To sum up, the reason for our exploration is to understand the seasonal aspect of data, acknowledging how wine theory correlates with the way data is distributed. Besides, level of glucose values and pH values are seen to vary by the period that samples were taken.

QUERY BUILDING

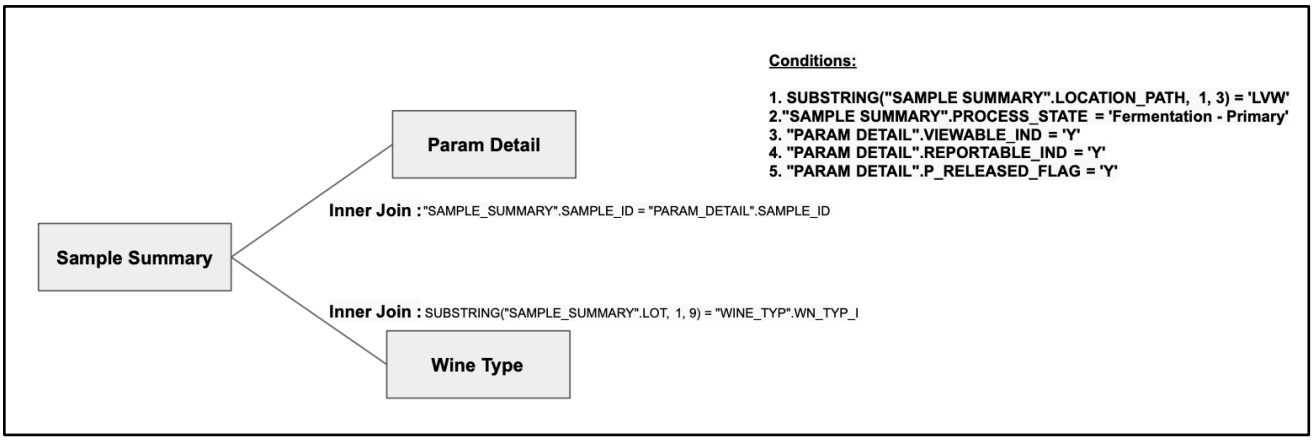


Figure D

Figure D illustrates the query structure implemented to retrieve the data for the first initial dataset containing glucose/fructose, NOPA, ammonia, PH values, total SO2, free SO2, Lactic acid and Malic acid. We joined the Sample Summary table with Param Detail tables with SAMPLE_ID being the common column to link these two tables together. In this multiple joins query, the Sample Summary table is also merged with the Wine Type table using the common column WINE_TYP. Then, we filtered the location to be Livingston for the site, and process state would be Fermentation-Primary. Lastly, in the Param Detail table, we also filtered the VIEWABLE_IND, REPORTABLE_IND and RELEASE_FLAG columns to be YES.

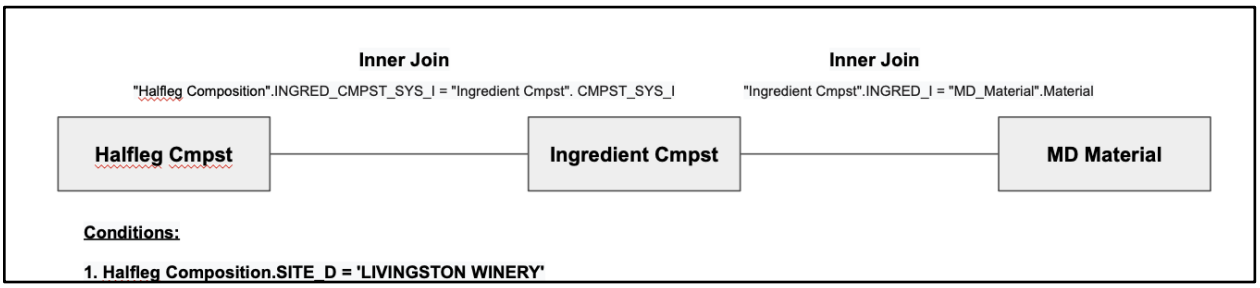


Figure E

Figure E provides us with details about the second query architecture we used to retrieve ingredient composition data. Particularly, we joined the Half Leg Cmpst table to Ingredient Cmpst table with common column as CMPST_SYS_I. In this multiple joins query, Ingredient Cmpst is also merged with MD_Material table with common column being Ingridient_I which is equivalent to Material column in Material table. The condition that we used for this query is to filter the site to be restricted to Livingston as the location.

Initially, we filtered the query with the condition Created Date restricted to the year of 2020 after we realized the significant size of the ingredient data. Afterward, we removed that filter to extract a larger range of data to be merged with more rows in our initial dataset.

DATA PREPROCESSING

For data cleaning and merging, our first step was to get separate extracts of glucose, NOPA, ammonia, PH values, total SO2, free SO2, Lactic acid and Malic acid into a csv file format from HANA. We used the inner join mechanism to combine all csv files into a final dataset and converted all p-value columns from string to numeric. This was done specifically to remove values such as '<0.5' to '0.5'. As Gallo's data had no column that would specify which wine category (white, red, sparkling, etc.) the data belonged to, we extracted the third number from the LOT column and used it to create a WINE_TYPE column based on the information from Gallo (0 represents red wine, 1 represents white wine). For example, if the LOT number is **WIN051WAS-18-L002**, the first number of the first 9 digits is 0, which tells us that this LOT number represents a red wine. Otherwise, if the LOT number is **SKC107CAL-20-L017**, the first number of the first 9 digits is 1, which tells us that this LOT number represents white wine. If the first three numbers are 101, it could also tell us that this wine is chardonnay. In the end we scratched the LOT column and decided the wine type in accordance with the first number in the initial 9 digits after the characters of the corresponding LOT.

Afterwards, we created an ingredient composition query to filter important columns for the ingredients composition, stated earlier, then combined it to the first datasets (glucose/fructose, NOPA, ammonia, PH values, etc.) using the common columns being EQUIPMENT_ID and LOT_I.. Lastly, we grouped the new joined dataset by TNK_I, LOT_I, S_CREATED_DATE, CMPLT_T and filtered only rows that have the latest completed time of work orders that happened right before sampling time of specific tanks and lots.

For missing values, we used python code to fill the null values based on existing values in the corresponding columns. The data is structured as a time series so gaps in the data can be filled by propagating the non-null values forward and backward.

GRAPH DATABASE

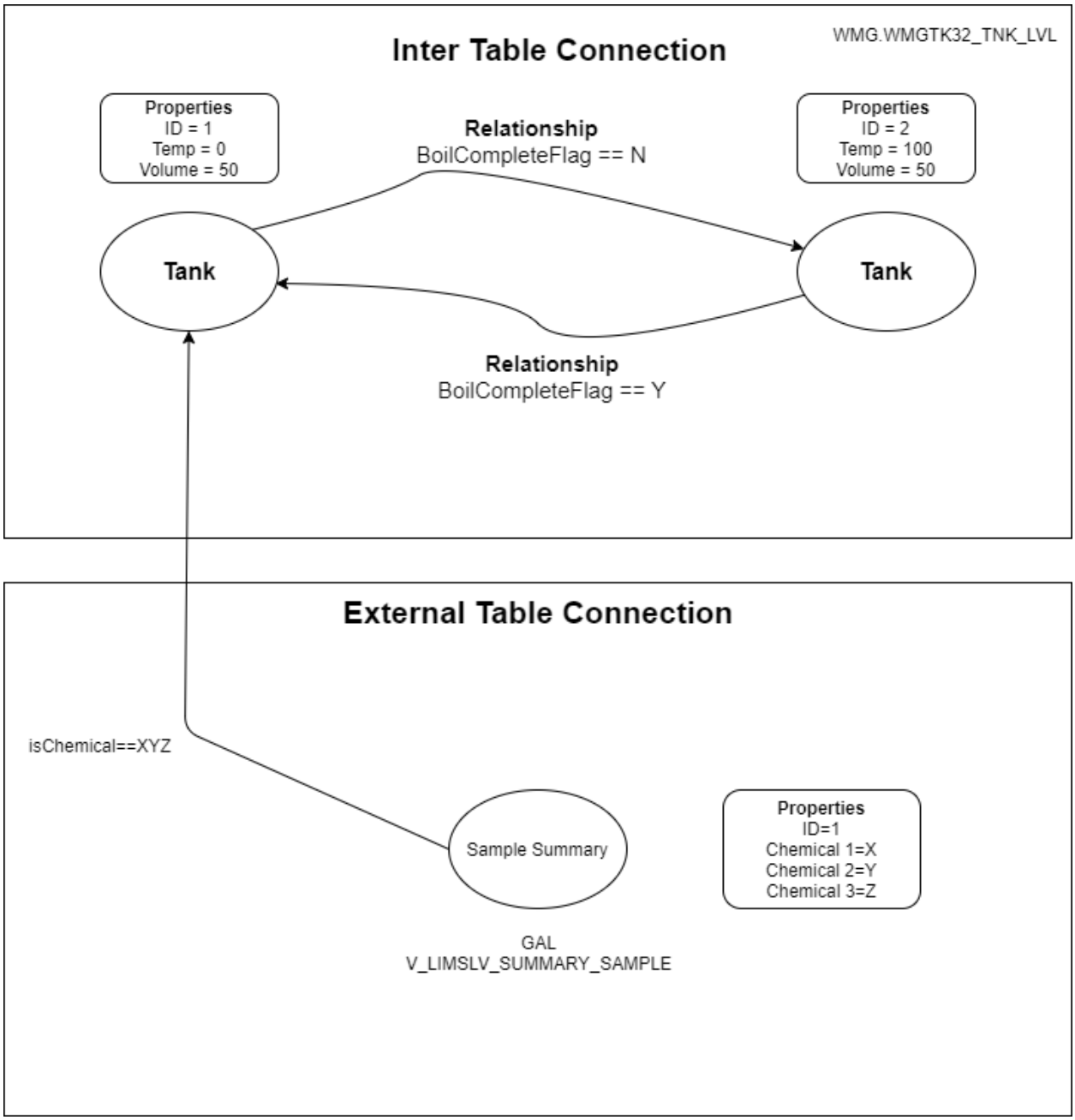


Figure F

We were initially instructed to build a graph database architecture for Gallo's tank data. The purpose behind building this was to keep track of which wines were being sent to which tanks. We developed and presented a basic architecture for this as shown in Figure F.

In our architecture we were treating each tank as a node, the actions as relationships between them and the measurements collected from the tanks as their properties. In the case above we gave the example

of BoilCompleteFlag, if it was set to ‘Y’ then the tank material will be sent over to another tank for the next process in wine making. We chose a process flag as a relationship because it would make it easier to track the progress of a wine in the making. We also had to provide a way for them to link up their sample data with their tank in future. For that we chose to put each sample as a node, chemical flag as a relationship and sample measurements like PH, Glucose/Fructose values as properties. However, we later changed the relationship from chemical to Tank IDs Gallo was not collecting chemical composition information when sampling. This also helped us to tie the sample to which tank they were taken from.

Nevertheless, we never got to implement it because Gallo’s Neo4j setup could not be configured for us to access, so that we could not develop and deploy the proposed architecture on and store the data accordingly. We were instructed to stop working on it mid-March.

PREDICTIVE ANALYSIS

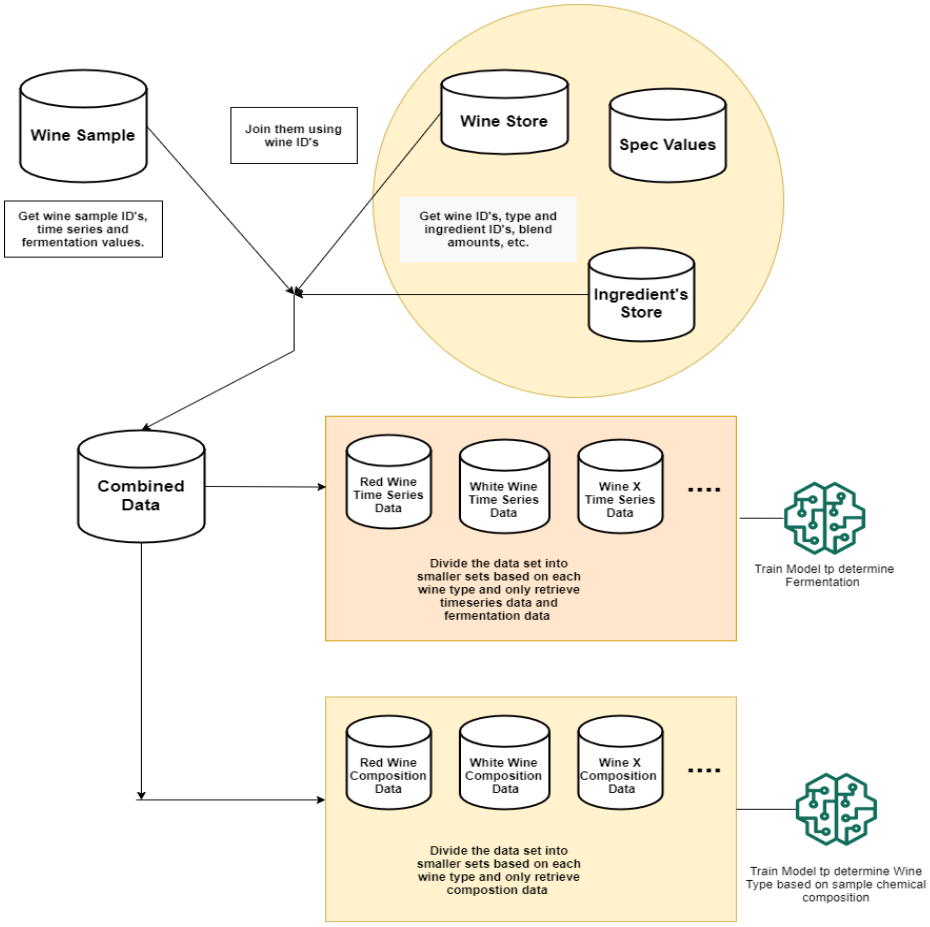


Figure G

Initially, we planned on doing time series analysis and forecasting. We planned on designing an app that would predict the fermentation stage of a certain wine type whenever its sample is taken out in the field using a two-step process.

In the first step we would have used a model trained on wine's chemical composition data, which will determine the wine type. We involved this first step to automate wine type detection as we were informed by the team at Gallo that lab data entry was done by hand and was prone to errors which we later saw when we received data access. The predicted wine type would be used to select the wine's specific fermentation prediction model as the second step. This model will be trained on fermentation data from wine samples and tank data.

When we got full data access and relative freedom to inspect Gallo's data we found that the designed strategy could not be used, due to many reasons. Firstly, there were too few wine samples taken, 296 to be exact in total for all types of wines. Secondly, the samples were taken at random intervals as seen

in Figure C. Lastly, the data was not stored accurately such as the column which stored all the sample measurements was not created properly and could accept string values such as '<0.5', could be left as blank, etc. Seeing all these issues, we determined that time series forecasting was not possible because one of the key things required to forecast is that data should be consistent and be in large quantities.

Therefore, we chose to go with regression analysis. We focused on just predicting fermentation (glucose/fructose) values. We chose regression analysis because we wanted to determine which factors were important to predict fermentation and to predict the value of it as well. We trained the extracted data over thirty plus different regression algorithms. The primary reason was to check which algorithm gives us the best results and then spend time on fine tuning its parameters at a later stage.

Model	Adjusted R-Squared	R-Squared	RMSE
ExtraTreesRegressor	0.73	0.75	4.97
XGBRegressor	0.69	0.71	5.35
RandomForestRegressor	0.66	0.68	5.58
KNeighborsRegressor	0.63	0.66	5.82
AdaBoostRegressor	0.62	0.64	5.92
DecisionTreeRegressor	0.62	0.64	5.95
BaggingRegressor	0.61	0.64	5.98
GradientBoostingRegressor	0.56	0.59	6.35
HistGradientBoostingRegressor	0.54	0.57	6.53
LGBMRegressor	0.51	0.55	6.68
SVR	0.36	0.41	7.64
NuSVR	0.33	0.38	7.82
Lars	0.31	0.36	7.96
TransformedTargetRegressor	0.31	0.36	7.96
LinearRegression	0.31	0.36	7.96
Ridge	0.31	0.36	7.96
SGDRegressor	0.31	0.36	7.96
BayesianRidge	0.31	0.36	7.97
OrthogonalMatchingPursuitCV	0.31	0.36	7.97
RidgeCV	0.31	0.35	7.98
ElasticNetCV	0.30	0.34	8.04
LassoLarsIC	0.29	0.34	8.05
LassoCV	0.29	0.34	8.06
LassoLarsCV	0.29	0.33	8.11
LarsCV	0.29	0.33	8.11
PoissonRegressor	0.27	0.32	8.17
LinearSVR	0.25	0.30	8.31
Lasso	0.25	0.30	8.33
ExtraTreeRegressor	0.24	0.29	8.34
HuberRegressor	0.22	0.27	8.47
GammaRegressor	0.22	0.27	8.49
ElasticNet	0.20	0.26	8.56
MLPRegressor	0.18	0.23	8.70
GeneralizedLinearRegressor	0.17	0.23	8.71
TweedieRegressor	0.17	0.23	8.71
OrthogonalMatchingPursuit	0.17	0.23	8.72
PassiveAggressiveRegressor	0.08	0.14	9.19
LassoLars	-0.08	-0.00	9.95
DummyRegressor	-0.08	-0.00	9.95
RANSACRegressor	-0.09	-0.02	10.03
KernelRidge	-2.08	-1.88	16.84
GaussianProcessRegressor	-17597.29	-16404.19	1272.14

Figure H

Model	Adjusted R-Squared	R-Squared	RMSE
ExtraTreesRegressor	0.83	0.93	2.56
DecisionTreeRegressor	0.77	0.91	2.98
XGBRegressor	0.73	0.89	3.19
RandomForestRegressor	0.73	0.89	3.22
GradientBoostingRegressor	0.65	0.86	3.66
BaggingRegressor	0.63	0.85	3.76
ExtraTreeRegressor	0.56	0.82	4.09
AdaBoostRegressor	0.47	0.78	4.51
KNeighborsRegressor	0.36	0.74	4.94
LGBMRegressor	0.17	0.66	5.64
HistGradientBoostingRegressor	0.16	0.66	5.66
LassoLarsIC	-0.58	0.36	7.77
SGDRegressor	-0.61	0.34	7.85
BayesianRidge	-0.62	0.34	7.86
ElasticNetCV	-0.64	0.33	7.91
RidgeCV	-0.64	0.33	7.92
LassoCV	-0.65	0.33	7.93
LarsCV	-0.67	0.32	7.98
LassoLarsCV	-0.69	0.31	8.03
Lasso	-0.71	0.30	8.08
OrthogonalMatchingPursuit	-0.72	0.30	8.11
TweedieRegressor	-0.74	0.29	8.16
GeneralizedLinearRegressor	-0.74	0.29	8.16
Ridge	-0.76	0.28	8.21
NuSVR	-0.79	0.27	8.27
ElasticNet	-0.80	0.27	8.28
SVR	-0.80	0.27	8.29
OrthogonalMatchingPursuitCV	-0.81	0.26	8.31
PoissonRegressor	-0.82	0.26	8.34
LinearRegression	-0.89	0.23	8.51
TransformedTargetRegressor	-0.89	0.23	8.51
MLPRegressor	-0.96	0.20	8.65
GammaRegressor	-0.99	0.19	8.72
HuberRegressor	-1.17	0.12	9.11
LinearSVR	-1.21	0.10	9.19
Lars	-1.24	0.09	9.25
RANSACRegressor	-1.32	0.06	9.42
LassoLars	-1.46	-0.00	9.69
DummyRegressor	-1.46	-0.00	9.69
PassiveAggressiveRegressor	-2.80	-0.54	12.04
KernelRidge	-5.76	-1.75	16.07
GaussianProcessRegressor	-1068.80	-434.17	202.14

Figure I

In order to establish a baseline, we first used the initial data set. It had just three features, collected as measurements when wine samples were taken out in the field. The results for the initial data set showed an adjusted R-Squared of 0.73 and RMSE of 4.97 results can be seen in Figure H. An adjusted R-squared value above 0.75 is considered good and shows that the features used were adding value to the model but the RMSE showed that there was still a significant margin of error. With the purpose of improving upon the model, we added the wine ingredient and composition data which increased the adjusted R-Squared

value to 0.83 and decreased the RMSE to 2.56. It gave us the confirmation that we are on the right track. Wine ingredients and composition data do add value to the model and also halved the previous RMSE. In both occasions (Figure H and I), ExtraTreesRegressor came back as the best performing algorithm, which is due to its extremely randomized nature and is known to work well with noisy data (Noisy data is data with a large amount of additional meaningless information in it called noise).

We conveyed our findings to Gallo and suggested they use ExtraTreesClassifier and the selected features in the final data set when they have gathered more wine sample data and are able to store internal wine tank temperature with the wine sample data.

CHALLENGES

To assist Gallo in moving forward towards successful modeling of wine fermentation, as well as to improve the process for future University of the Pacific Capstone teams, we have compiled a list of recommendations that emerged from the challenges we faced over the course of the project and our understanding of Gallo's data. Our most significant challenges arose from three distinct issues: data access, data complexity, and the quality and quantity of available data.

One of our biggest challenges was achieving full and consistent access to the data we required. It wasn't until five weeks into the project that any on our team were able to explore Gallo data firsthand. At that point some of the team gained access to Oracle data, but firewalls and VPN issues extended the delay into week 6 for Mac users. Then, slow export speeds made it impossible to work with large quantities of Oracle data, so Josh began working on getting us access to the HANA database, which we started during the eighth week of the project. We were never able to achieve working HANA credentials for the entire team; however, and retaining access to any of the databases continued to be a challenge throughout the project. An additional problem was with the completeness of the data we were using; on several occasions we discovered important data missing from HANA, further slowing our ability to understand and use the data effectively, especially given its complexity.

This data complexity was the second significant challenge we faced. We were required to form our understanding of the available data primarily through verbal communication and live demos without written documentation from a data dictionary or entity relationship diagram, and also though answers to specific questions were provided by emails or on Microsoft Teams. The same was true for information about Gallo's wine fermentation process. It was not until early March that we were given diagrams depicting Gallo's three distinct fermentation processes for white, red, and sparkling wine. One team member was eventually able to locate a Gallo sales manual from 2002 containing over 500 pages of detailed explanations useful for decoding portions of the available data, but the ad hoc nature of our access to this information slowed our progress significantly, as did our inability to meet with winemakers to gain domain-specific knowledge about the processes captured in Gallo's data, despite initial optimism about the possibility of doing so and several meetings scheduled for this purpose. While our Gallo contacts did their best to provide us with the information we needed, a deeper domain understanding would have helped us extract relevant data much more quickly and effectively, and thus contribute to the creation of a more effective model.

Because of the delays arising from these challenges, it was not until the 13th week of the project—two weeks before the final presentation to Gallo—that we were finally able to extract the necessary data

and began quickly merging and cleaning the data for model use. In the end, the data that we had was not sufficient to predict the best time to bottle Gallo wine well. As our team was unable to extract tank temperatures that corresponded to lab analysis of juice undergoing fermentation, we were missing one of the most important features. The small number of samples (296 rows) available for modeling, and the inconsistent timing of sampling, were additional issues. Thus, we needed more features to make a better model with higher accuracy.

RECOMMENDATIONS

Although Gallo has deep data going back many, many years, it appears based on our work that additional data of several types is needed. Thus, the first and most important recommendation is for additional data collection. Specifically, improved collection and storage of tank temperature is important, due to the role of temperature in fermentation. When storing tank temperature, we recommend that local temperature also be stored due to its potential importance as a factor. Further, if samples are taken at a set schedule and consistently, then a forecasting model could be created using deep learning to build an RNN—or even a simple time series—to predict fermentation values at a given time for specific wines.

Therefore, as immediate next steps towards achieving successful modeling of fermentation, we recommend Gallo collect three types of additional data. First, Gallo should collect tank and local temperatures at regular and frequent intervals; we suggest 5 minute intervals. Second, Gallo should ensure that samples of fermenting juice are taken on a set schedule and at least three times a day for a minimum of one month. Third, collecting a "finished" timestamp when fermentation is complete is also useful. It's critical that this data be high quality—that is, be consistent, accurate, and complete, so we recommend that a data quality engineer oversee its collection. Successful collection of this additional data—even for a limited period of time—would make it possible to use statistical models and mathematical equations to simulate data to build predictive models in the short term.

Our second recommendation is that Gallo provide earlier, more consistent, and more complete access to the data to allow for faster identification of problems with data quality and quantity (such as those our team encountered when trying to extract tank temperature and other features), as well as earlier identification of knowledge gaps so that they can be addressed through independent research or consultation with appropriate Gallo stakeholders. In this way, more time would be available fine-tuning feature extraction and improving model performance.

Our third recommendation is that written documentation relevant to the project be provided to team members as part of the initial orientation. A data dictionary and entity relationship diagram would be especially helpful, and would be useful to regular Gallo employees involved in data analysis and manipulation as well. We also recommend including Gallo-specific information relevant to the project (like the diagrams and sales manual mentioned previously). Access to relevant documentation would also reduce the need for extended consultation with Gallo staff throughout the project, thus reducing the strain of integrating an additional project into existing work responsibilities.

Lastly, our final recommendation is to increase access to domain expertise by involving key stakeholders—like Gallo winemakers—in the project. This would help future teams deliver better business value through increased understanding of stakeholders' needs early on in the project. It is especially

important to be able to consult internal domain experts when outside help is not possible because of non-disclosure agreements.

In overall, we understand that the ability of any organization (including Gallo) to provide frictionless access for a project of this type can be limited for a wide variety of reasons, ranging from time constraints and security concerns to hacker attacks and a global pandemic, and we are deeply appreciative of the opportunity to work the Gallo to use complex data to attempt to solve a real world problem and create business values. It is our hope that the lessons learned from the challenges we faced over the course of the project can prove valuable to all involved, contributing to increased success for future projects.

ACKNOWLEDGMENTS

We would like to thank the following individuals for their support and without whom completion of this project would not have been possible: from E. & J. Gallo, Joshua Shackleford - DevOps Manager, John Reinhardt - IT Specialist, and Jessica Parsons - Business Analyst; from the University of the Pacific, our advisor, Dr. Emma Hayes, and program directors Dr. Rick Hutley and Dr. Jim Hetrick.

REFERENCES

Henschen, D. (2012, September 5). *How Gallo brings analytics into the winemaking craft.* InformationWeek. https://www.informationweek.com/it-leadership/how-gallo-brings- analytics-into-the-winemaking-craft/d/d-id/1106133

IBM Research Blog (2017, April 6). *From IoT and vines grow the fruits of innovation.* https://www.ibm.com/blogs/research/2017/04/iot-grows-innovation/

Jamieson, B. (2007, March 8). *Ernest Gallo, the truth behind the myth.* ABC News. https://abcnews.go.com/Business/story?id=2934757&page=1

McGowan, B. (2017, July 18). *E. & J Gallo deploys data-driven app to harvest a better grape.* CIO. https://www.cio.com/article/3209088/e-j-gallo-deploys-data-driven-app-to-harvest-a-better-grape