Wine Weather: Exploring Meteorology and Grape Harvests in California

Mariana Gurgel, Rohan Devulapally, William Ruan, Subbi Sathya, Ruhaan Shinde, Minchae Yun, Justin Zhao

PURDUE UNIVERSITY | The Data Mine

BASF We create chemistry

INTRODUCTION

BASF is a leading chemical company with the goal to combine economic success with environmental protection and sustainability. With that in mind, our goal was to develop a machine learning model that could predict future grape yield using weather forecast data in California. With this model, we can help farmers optimize their wine production by giving them an estimate of future yield that allows for better allocation of resources.

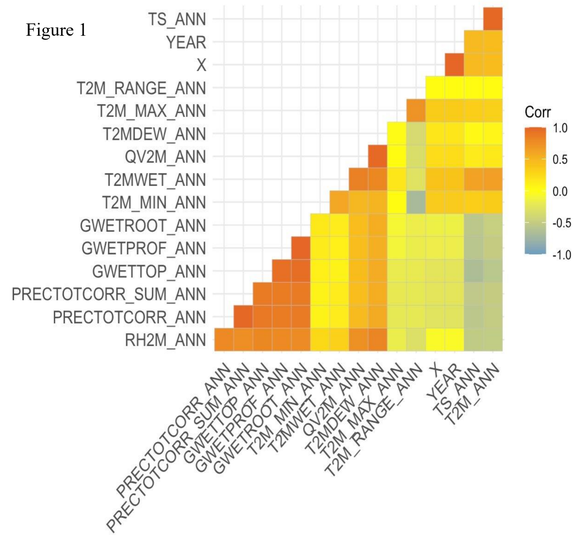
Objectives:

- Learn about weather patterns and their effect on grape yield
- Create a functional machine learning model capable of predicting future grape yield based on weather forecast data
- Use hyperparameter optimization to fine tune our machine learning algorithm
- Provide this data to farmers through a user interface

Exploratory Data Analysis

Figure 1 is a correlation table that uses r values to look at how some of our variables correlated to each other. The darker areas in the graph show a stronger correlation.

Figure 1



Data Descriptions

Nasa Power View Weather Data
The NASA weather viewer allows us to see weather data for a target locations given longitude, latitude, and dates.

Kaggle California
Contains up to 40 years of grape wine production data from California that was used for labelling data.

Daily Datasets
We wanted to use daily data to increase the accuracy of our model. For that purpose, we had to merge our daily dataset to our yearly yield dataset. This led to issues when merging data frames of different sizes. Figure 2 shows how we fixed these issues by transforming our daily data into arrays and rearranging them into a single row for each year of data composed of all their daily variables.

4741	4742	4743	4744	YEAR	PRICE(Dollars/Ton)	YIELD(Ton/Acre)
8.98	10.30	16.32	6.01	1981	404.00	4.07
2.33	14.20	12.17	-2.03	1982	249.00	3.41
11.78	10.71	20.31	9.60	1983	275.00	4.12
7.28	14.35	17.34	2.99	1984	335.00	3.90
12.60	8.82	19.62	10.79	1985	397.00	3.71
9.21	12.02	20.07	8.05	1986	393.00	4.41
3.32	11.98	11.29	-0.69	1987	407.00	3.94
3.33	14.47	13.12	-1.35	1988	537.00	3.32

REFERENCES & ACKNOWLEDGEMENTS

- NASA Power View Weather Data: https://power.larc.nasa.gov/data-access-viewer/
- Kaggle - "California Wine Production 1980-2020": https://www.kaggle.com/datasets/jarredpriester/california-wine-production-19802020
- Nappa Valley Vinters - "The Life Cycle of A Grape": https://napavintners.com/napa_valley/life_cycle_of_a_grape.asp

- Tracy Rowlandson - BASF Technical Marketing Group Manager
- Jason Chia - Teaching Assistant
- Emily Hoeing - TDM Corporate Partners Advisor
- David Glass - TDM Managing Director
- Cai Chen: TDM Corporate Partners Technical Specialist

METHODOLOGY

Hyperparameter Tuning

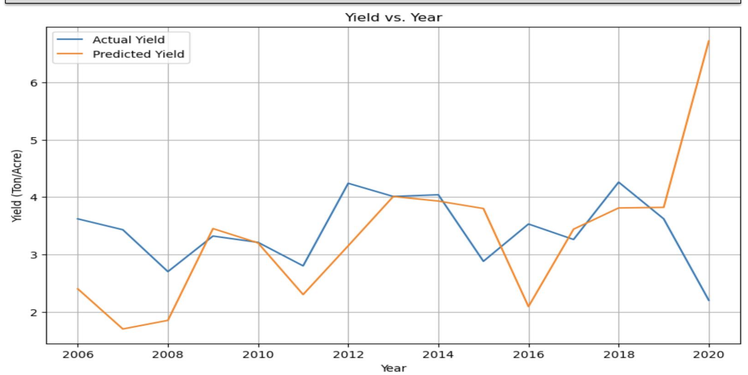
- Hyperparameter tuning experiments with different external variables to improve the accuracy of our machine learning model predictions, focusing on grape yield forecast.
- By implementing grid search, testing various hyperparameter combinations, we can evaluate their effectiveness across different parameter ranges to optimize model performance.
- The objective is to identify the hyperparameter set that minimizes loss, thereby enhancing the alignment between predicted and actual grape yields, ensuring the most effective model training and outcome.

Same Year Prediction Models

Allow us to look at variables and data from a specific year and use that information to predict yield of the same year. In these models no future prediction is being made, instead we are predicting values of the same year we are analyzing data for.

Time Series Ridge Regression Model:
This was the initial model we started developing based on weather forecasting models that utilize machine learning and ridge regression to predict future weather. We updated and modified the model to predict yield based on our most correlated variables like temperature, precipitation, humidity, and more.

LSTM Model
We created an LSTM model that successfully stores information on all the datasets for all of the counties, based on the daily data. We worked closely with the daily data team in order to organize their data and store it in one place that is accessible to all. The model was built on a previous model made for past data, and modified for the current data and also takes account for future years and leap years.



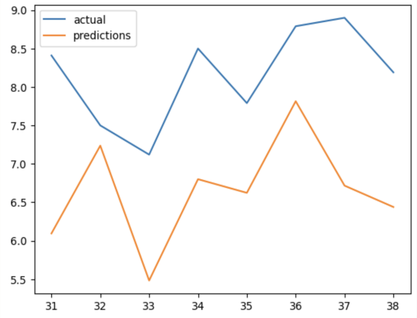
Graph of the LSTM output versus actual label data from 2006-2020.

UI + MODEL

We created a GUI to implement our code through a user interface for nonprogrammers.

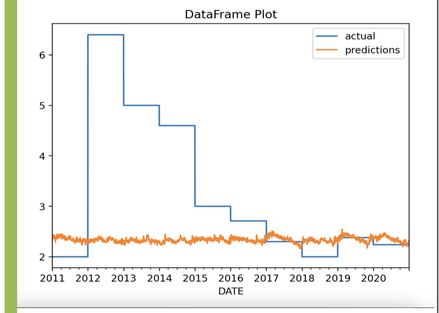
The UI consists of the following features:

- Box to drag and drop or upload files consisting of the finished daily data
- Upon inserting a file, the UI will generate a preview of the data, so the user knows the general logistics
- Furthermore, the UI also generates a graph based on the models we created



Graph of the ridge model output versus actual label data.

Grape Yield Model



Example graph output from the user interface output over held out test data across 2011-2020.

CONCLUSION

Graphic:

- For the graphic, we made a user interface that accepts any downloaded daily dataset on the sidebar
- The user may select what aspect of the data they would like to see. This includes statistics of the set as whole, a head of our newly formed daily data, and a graph of an LSTM model for the yield prediction.

Daily Data:

- For the Daily Data, we were able to finalize converting the yearly data to create a corresponding clean labelled daily format.

Hyperparameter Tuning:

- Implemented external variables and grid-search to optimize settings for the machine learning tests on the data.

LSTM Model:

- Training and testing the LSTM off of past data yields to predict future yields, we modified the parameters to maximize accuracy on predicted vs actual results and obtained some high percentages.

FUTURE GOALS

Daily Data

- Update the LSTM to make the daily data compatible with a working model
- Bug fix the model and make it easy to access all the data in a secure and readable location
- Update the data to ensure the model can predict daily changes given each day's yield being the same value for each year

Graphic

- Make the graphic function on a permanent network rather than making the user boot it up every time
- Add more features such as a dropdown for counties and a section to display results in a simple, written out format

Time Series

- Making a better, more accurate model for forecasting, as well as better detection for anomalies and outliers
- Make the analysis function in real time without much delay for optimal efficiency
- Make the hyperparameter tuning LSTM compatible with each county to ensure each dataset is accounted for and so overall trends can be inferred