Assortment Analysis

Sauvignon Blanc

TW_SauvignonBlanc_size6

Total Wine & MORE

Meet The Team!

Amlan Mohanty

Anwesha Mohanty

Atharva Tilak

Navya Gehlot

Rajat Marathe

Siddhant Soymon

Total Wine & MORE

Executive Summary

Total Wine & More currently performs demand forecasting manually using Excel-based models, which can be time-consuming and limited in accuracy. Our project developed a Machine Learning-based solution to enhance forecasting efficiency and optimize product assortments across 269 stores.

We used three key data sources:

- **Internal Sales Data:** Sales and inventory over the past 52 weeks.
- **Store Data:** Demographics and store tiers.
- **External Sales Data:** Market-level sales and distribution points.

The final model, **XGBoost** , was selected for its strong performance, achieving **MAE: 1172** and **R^2: 84.7%**

This model provides accurate demand forecasts and data-driven recommendations for optimal product assortments at each store, improving sales and inventory management.

Total Wine & More

Table of Contents

Total Wine & MORE

01 | Business Objectives

Total Wine & MORE

Business Objective

Develop a forecasting model based solution to address the limitations of Total Wine's current Excel-based models, enhancing processing efficiency and predictive accuracy.

Leverage these demand forecasts to develop data-driven recommendations for optimal product assortments tailored to each store, driving sales growth and improving inventory management.

2 | Data Highlights

Total Wine & MORE

Data Source

Internal Sales Data

Data about the normalized sales of items at each store and the number of weeks the item was in stock over the last 52 weeks.

Store Data

Data about each of the 269 Total Wine stores, the demographics of people within 5 miles of each store, and the store tiers.

External Sales Data

Data about the total market sales of an item in each state and the number of points of distribution for the item.

Total Wine & MORE

Data Highlights

Out of 435 unique wines, 108 wines are categorized as vintage

Most wines sold are of Package type 750ml.

Prices range from $3.49 to $222

Store sizes are mostly Large or Medium and only one Extra Large store is found.

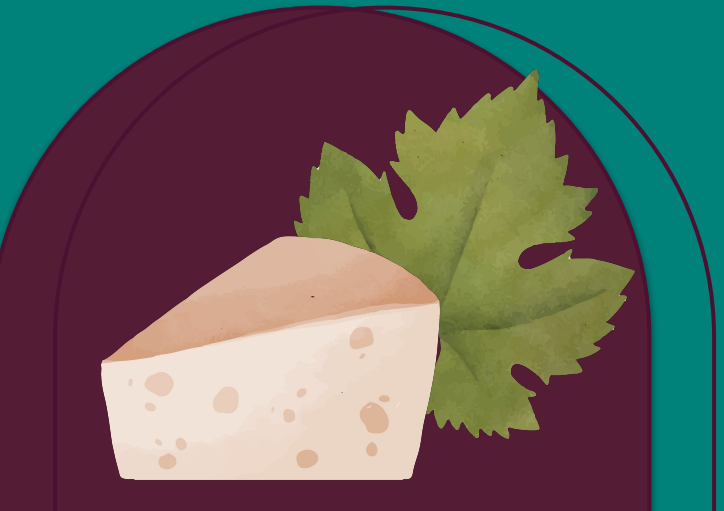
Total Wine & MORE

3

Data Cleaning & Feature Engineering

Total Wine & MORE

Data Cleaning



- Imputed the missing retail values of certain items
- Dropped rows whose item retail price could not be found
- Dropped certain irrelevant columns like store tier information for other alcohol categories
- Imputed missing values in the '*sales_dollars_L52wk*' and '*points_of_distribution_L52wk*' columns by using a state-wise average value. For states with no data on these columns, the average from similar states were imputed. (DE uses MD, KS uses MO)

Total Wine & MORE

Feature Engineering

01

Created a new column called '*Open Since Years*' which stores how many years since the store was opened.

02

Created a new column called '*is_vintage*' which stores whether or not an item is categorized as vintage.

03

Converted categorical variables to dummy variables

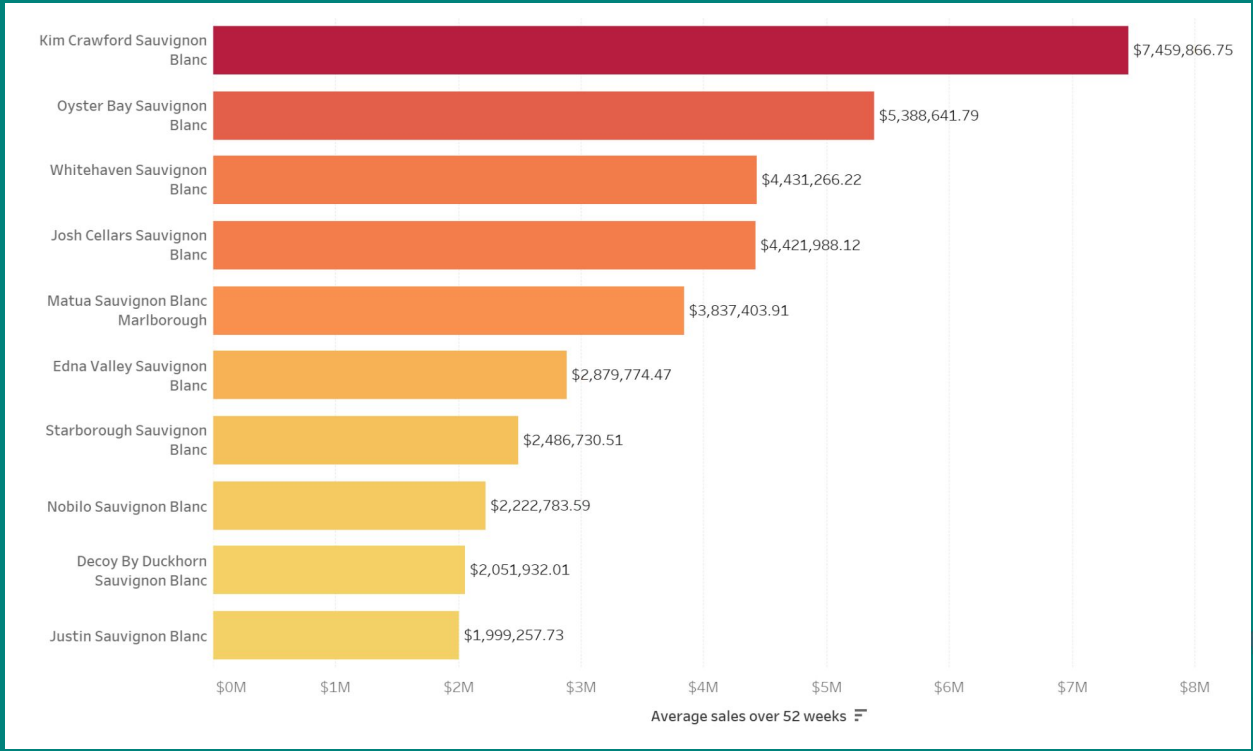
FEATURE SELECTION: We used the scikit-Learn's SelectKBest() method to select the best features for our models. SelectKBest uses statistical tests like the chi-squared test to score features based on their relationship with the target variable.

Total Wine & MORE

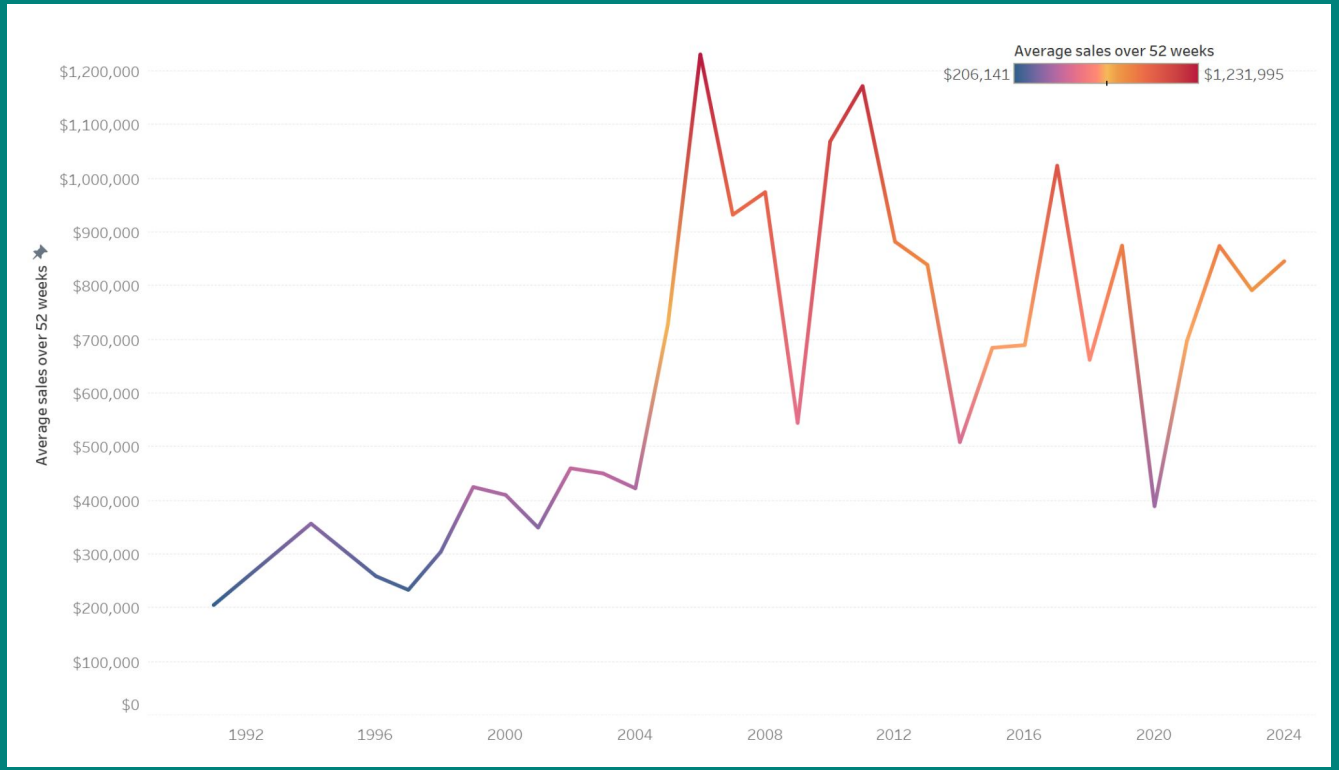
4 | Exploratory Data Analysis

Total Wine & MORE

Kim Crawford Sauvignon Blanc dominates the market, with sales nearly 40% higher than its closest competitor, Oyster Bay.

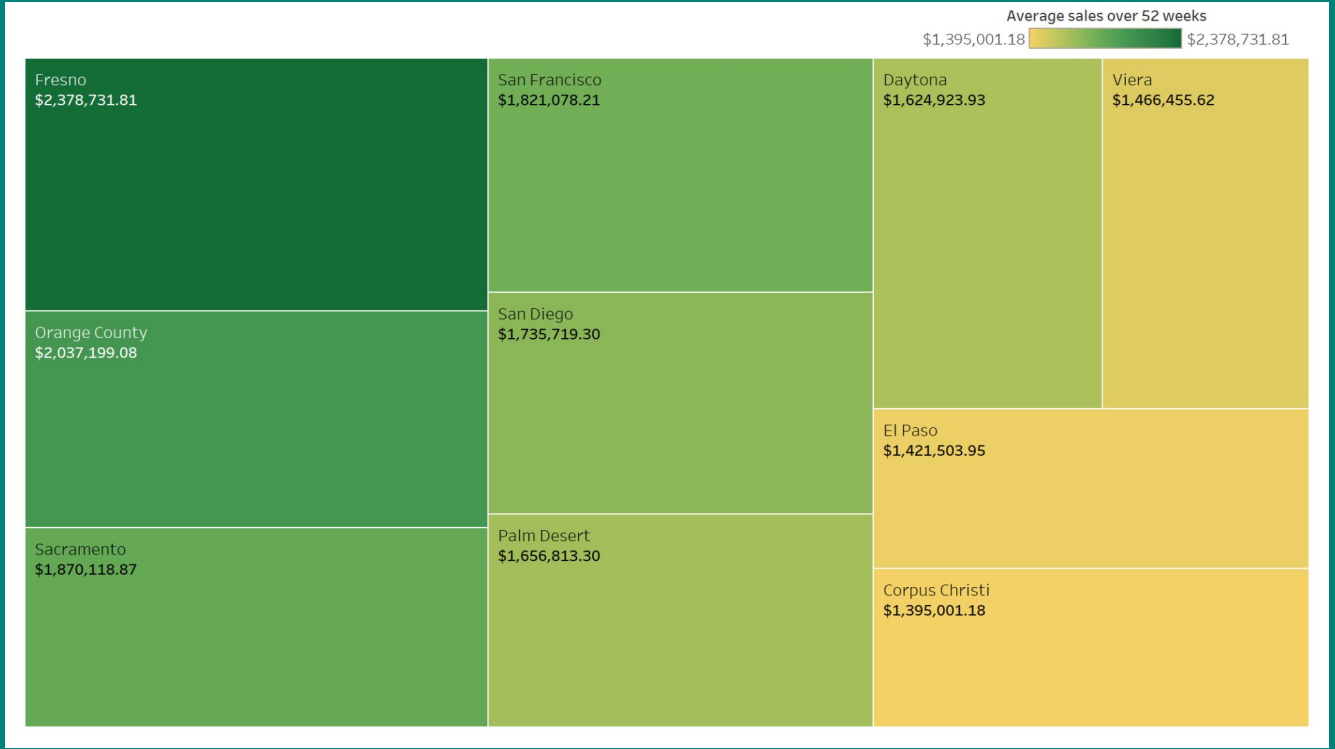


Average sales have shown significant volatility since 2008 but are now stabilizing near $1 million.

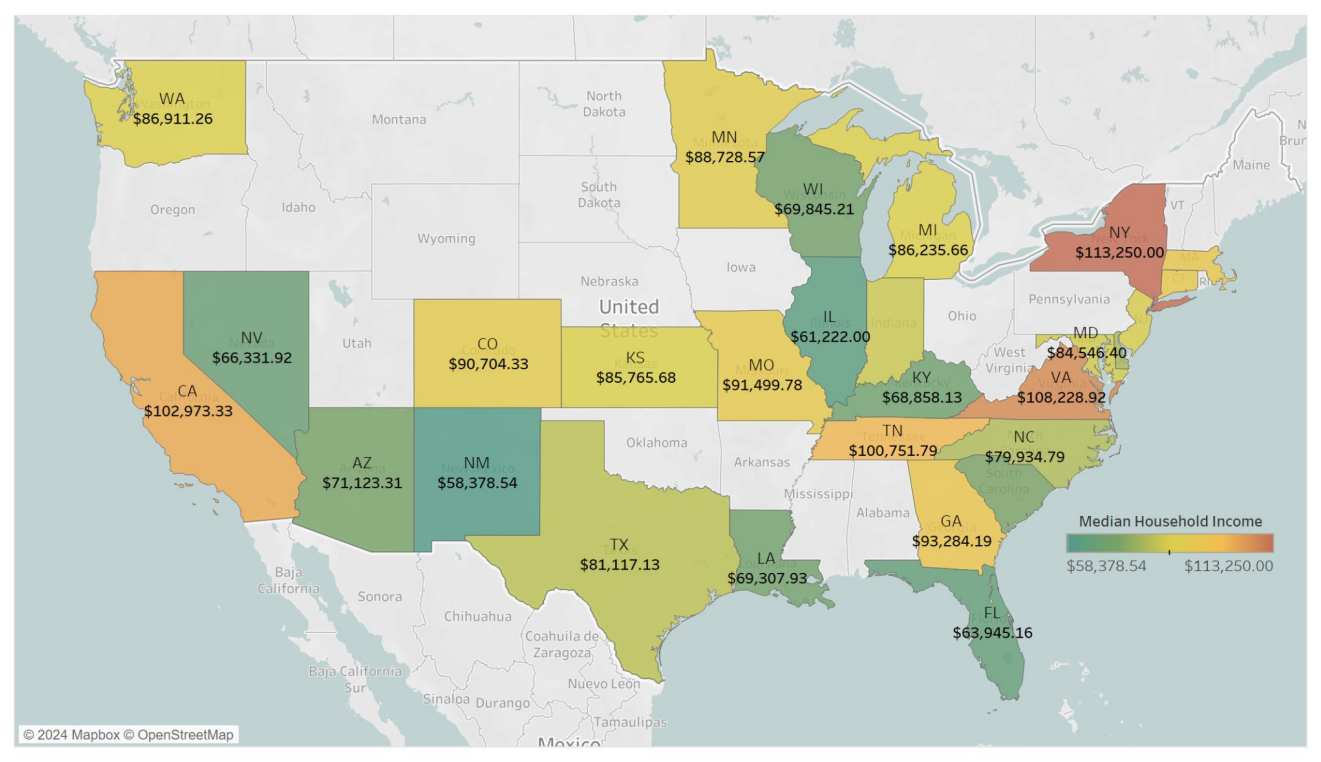


Total Wine & MORE

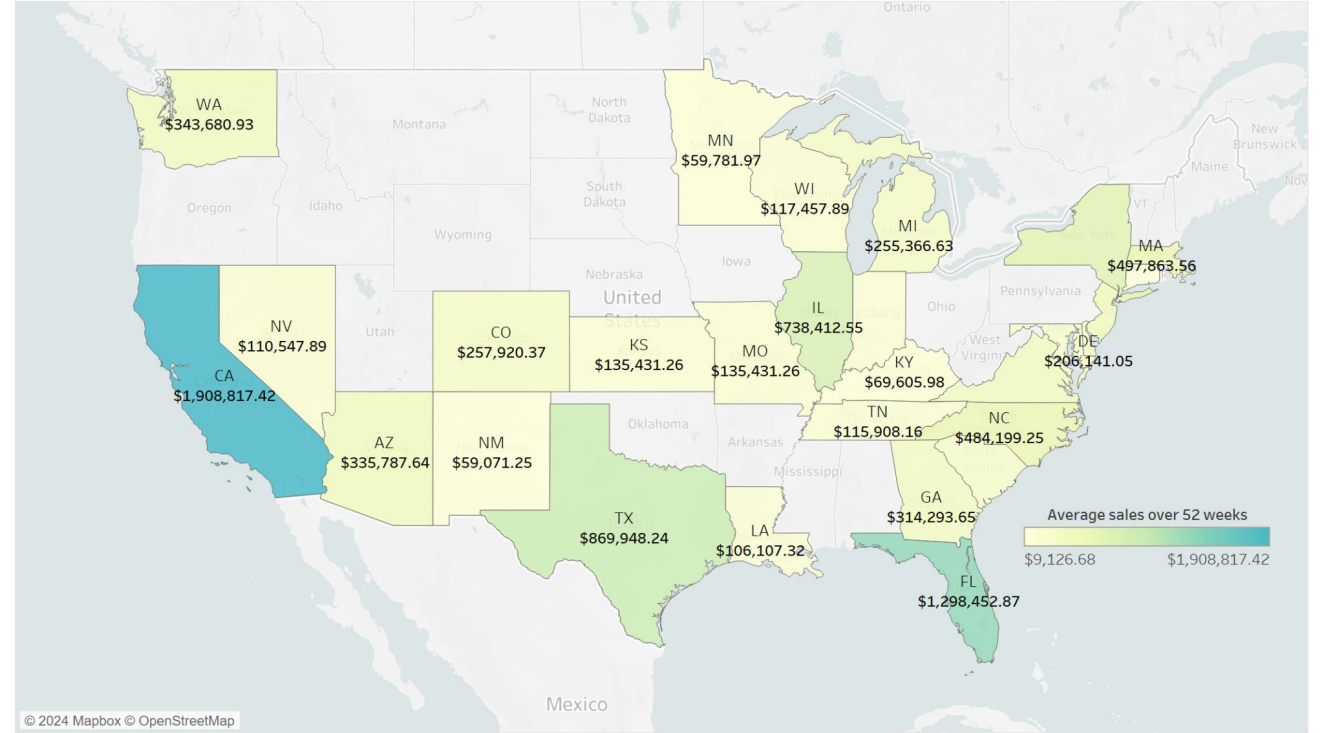
While Fresno leads in sales, Orange County and Sacramento contribute nearly as much as Fresno, highlighting the combined strength of these mid-performing markets.



Despite New York's high income, Southern and Midwestern states like Texas, North Carolina, and Louisiana show lower average incomes, highlighting regional economic disparities.



High-income states like New York and Virginia have lower sales than California and Florida, suggesting inefficiencies of market conditions.

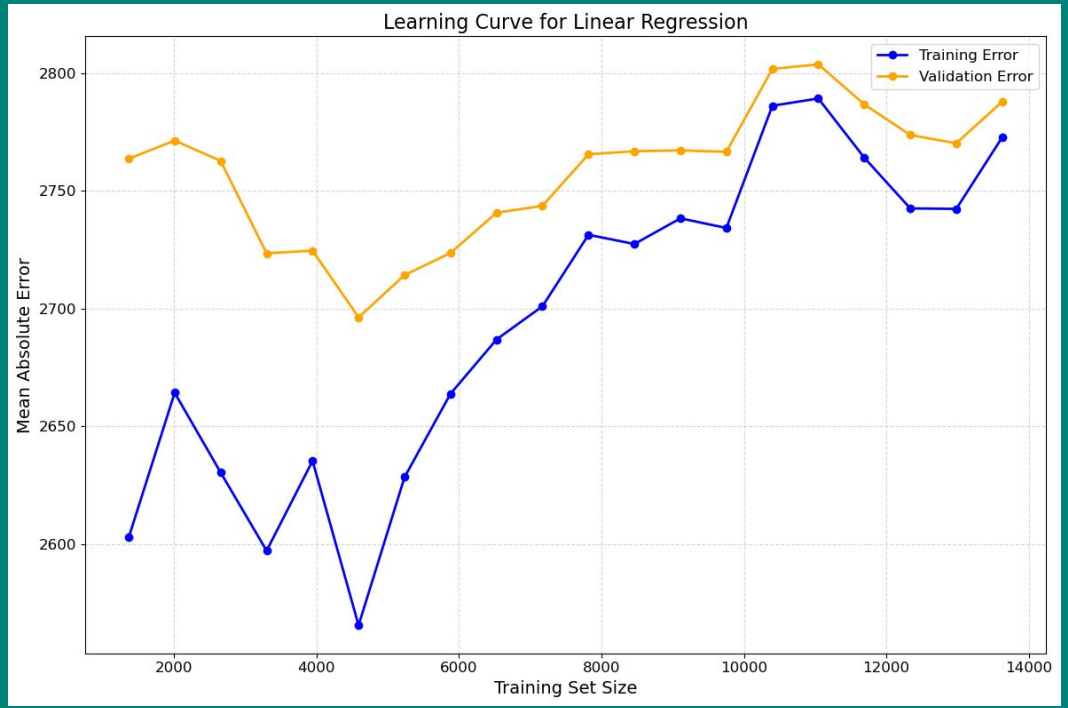


5

Model Description

Total Wine & MORE

Model 1 : Linear Regression



- Fits a linear equation to observed data
- Assumes linearity between dependent and independent variables

PROS

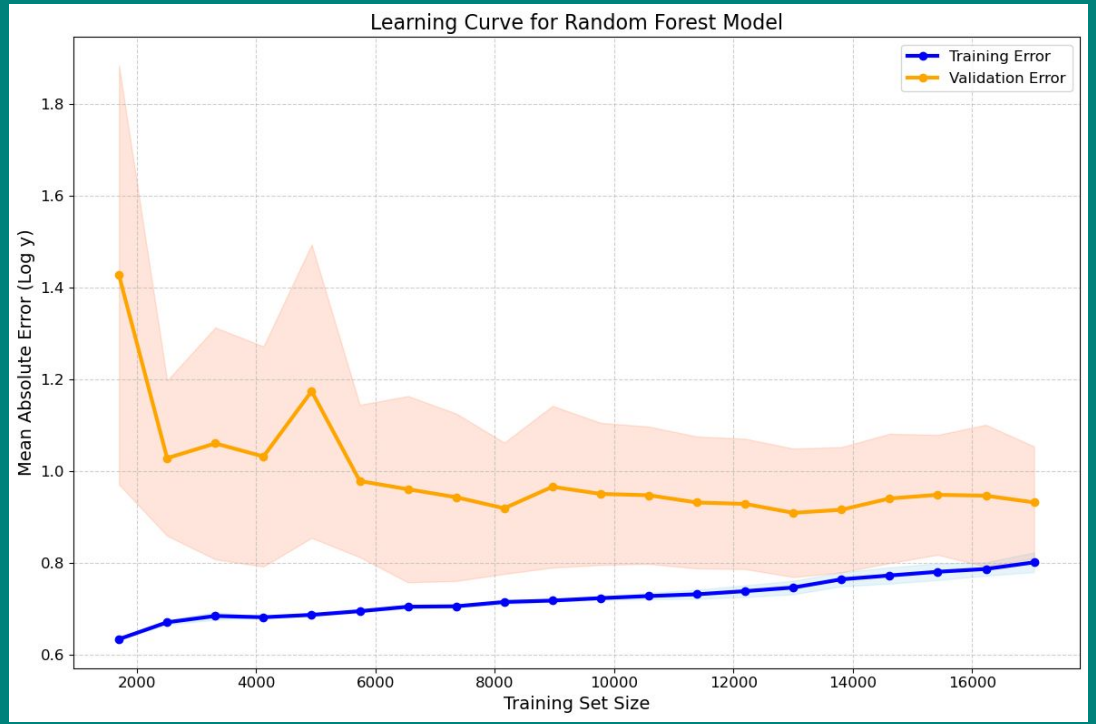
- Simple, interpretable, computationally efficient.
- Provides clear insights into the relationships between variables.

CONS

- Prone to the influence of outliers.
- High MAE (2752.89) and poor R^2 (35.71%) value.

Total Wine & MORE

Model 2 : Random Forest



- Ensemble learning method using multiple decision trees
- Combines tree predictions to improve accuracy

PROS

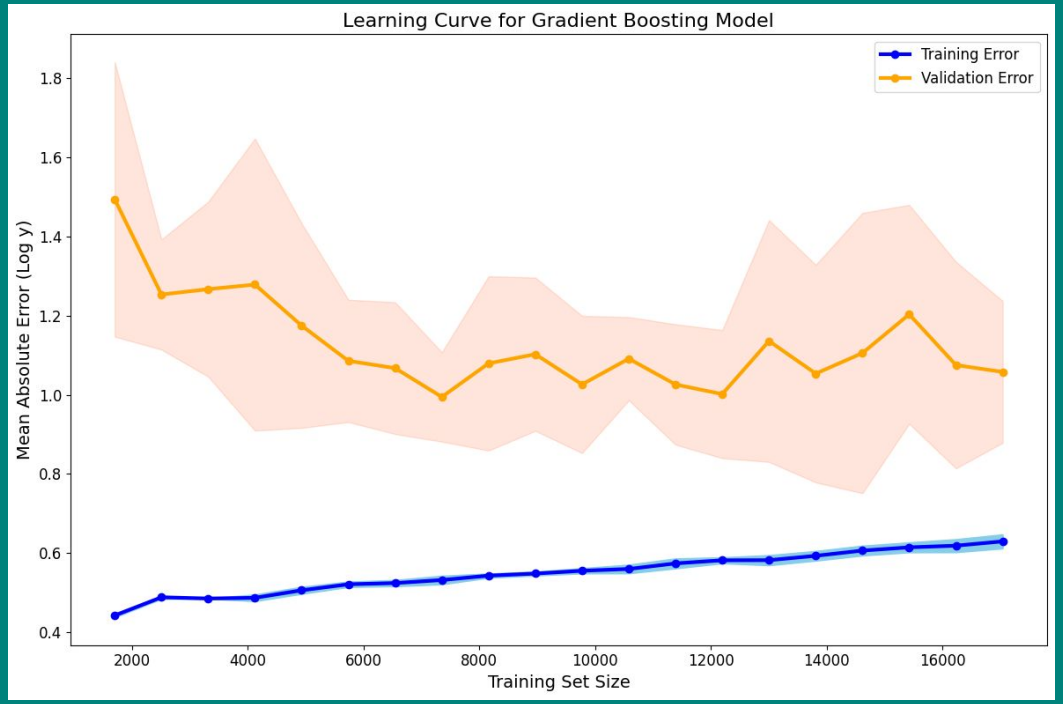
- Captures nonlinear and complex relationships better.
- MAE: 1956.95 (29% improvement)

CONS

- More complex, harder to interpret, slower to train.
- There's no significant difference in the R^2 values.

Total Wine & MORE

Model 3 : Gradient Boosting



- Ensemble boosting technique combining multiple weak models to create a strong predictive model.
- Learns from the errors at each iteration.

PROS

- MAE: 1307 R^2 : 80.48%
- High predictive accuracy
- Captures complex patterns in the data.
- Highly customizable with hyperparameters to tune.

CONS

- Longer training time.
- Prone to overfitting

Total Wine & MORE

Model 4 : XGBoost

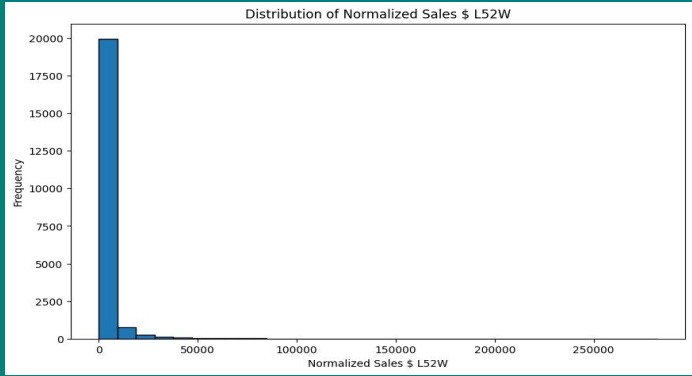
PROS

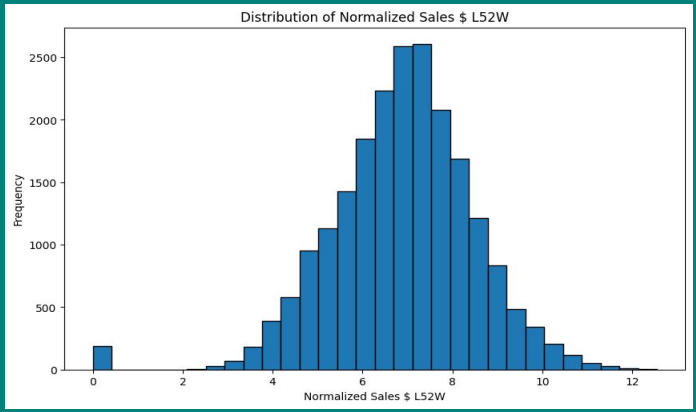
- MAE: 1266 R^2: 81.65%
- Highly accurate
- Faster training compared to gradient boosting.
- Robust with complex models.
- Our best model so far

XGBoost is an optimized, faster and more efficient version of gradient boosting algorithm.

Total Wine & MORE

Log Transform



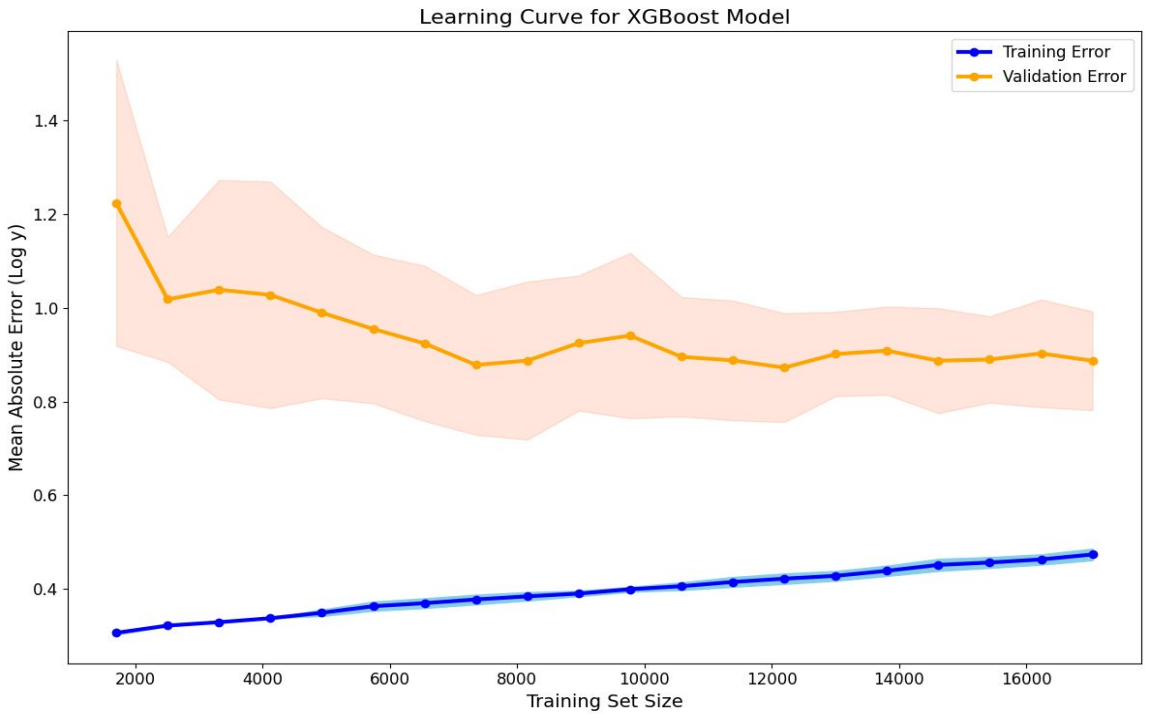


Hyperparameter Tuning

Parameters	Description
n_estimators	Number of trees built for boosting. Higher values mean more complexity.
max_depth	Controls maximum depth of each decision tree. Higher values mean more complex model.
learning_rate	Controls the step size at each iteration to control overfitting. Lower values mean more complex model.
subsample	Fraction of data to be randomly sampled during training.
colsample_bytree	Fraction of features to be randomly selected for each tree.
alpha	L1 regularization. Helps in preventing overfitting.

Total Wine & MORE

Model 5 : XGBoost (2nd version)



PROS

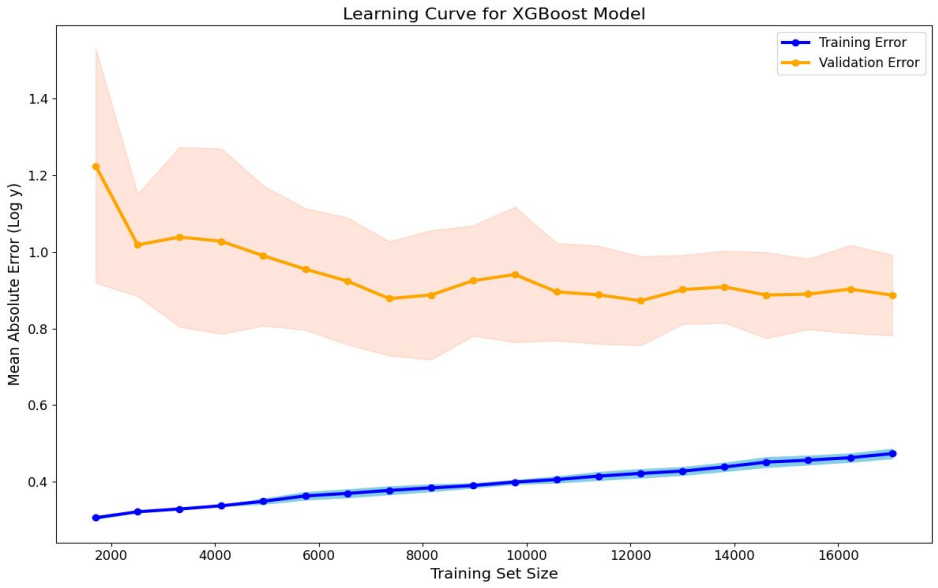
- Chose XGBoost as the final model after
  - Log Transformation of target variable
  - Hyperparameter Tuning using RandomizedSearch CV
- MAE: **1172** R^2: **84.7 %**
- Validated the results using K-fold cross-validation

6

Findings & Insights

Total Wine & MORE

Findings and Insights

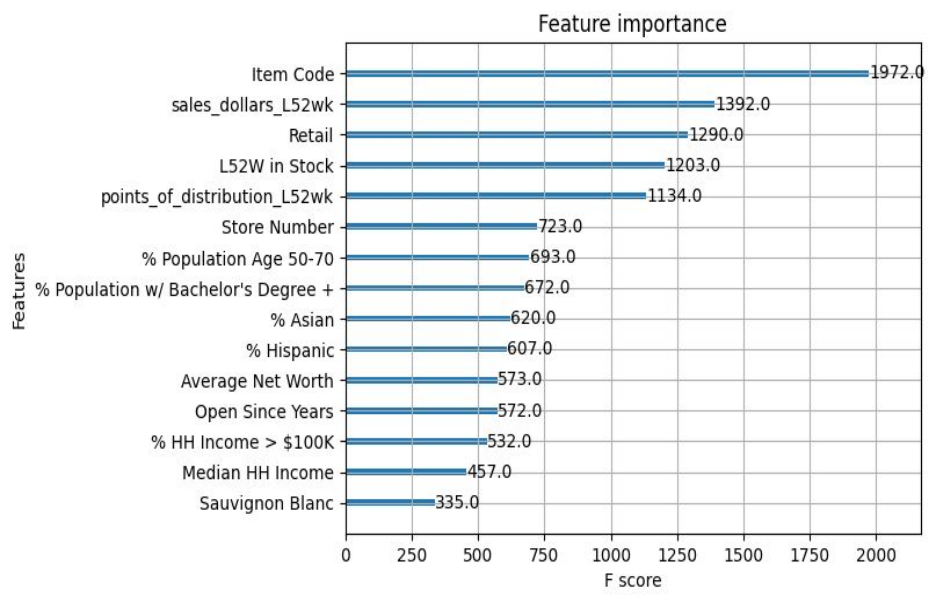


FINDINGS

- Predictive power increased significantly.
- High R^2 and low Mean Absolute Error.
- Works well with skewed data and handles outliers.
- Learning curve reveals decreased validation error with training size.
- Model utilizes available data effectively.

Total Wine & MORE

Findings and Insights



INSIGHTS

- Pricing strategies heavily influence sales.
- Larger points of distribution lead to greater product visibility.
- Stock availability is critical for driving sales.
- Age, Educational Qualifications and Cultural ethnicity can affect product preferences.
- Economic factors like Average Net worth and Household Income influence purchasing decisions.
- Number of years since a store has opened affects its popularity in the area.

7 | Challenges & Workarounds

Total Wine & MORE

Challenges & Workarounds

Highly Skewed Target Variable

Took Log Transformation for Target Variable for better model performance

Missing Values for Sales in External Data

Used interpolation for imputation, replacing missing values with data from similar states.

Total Wine & MORE

8 Recommendations & Opportunities

Total Wine & MORE

Recommendations

To develop a list of product recommendations that we should add to certain stores based on the likelihood of high sales, we followed the following steps -

- Develop a predictive model to predict the missing sales values of the test data
- Develop a list of all possible product and store combinations
- Find the closest store to every other store based on their characteristics
- Get the missing sales values for all possible store product combinations using the sales values(if they exist) for the closest store
- For the remaining missing values, replace them with average sales of the product in the state
- In case that product has never been sold in the state, replace them with the overall average sales of the product

Total Wine & MORE

Recommendations

- Consider only those store-product combinations that would likely have sales greater than $5000.
- Our final recommendation list includes a list of 10 products that should be introduced to 10 stores across 4 states.

Store Number	Item Code	Item Name	Store State	Predicted Sales
1502	4350750	Simi Sauvignon Blanc	CT	5487.00
302	231952750	Sunny with Chance of Flowers Sauv	NJ	5149.33
304	231952750	Sunny with Chance of Flowers Sauv	NJ	5149.33
301	231952750	Sunny with Chance of Flowers Sauv	NJ	5149.33
939	2343750	Cakebread Sauvignon Blanc	FL	5306.67
940	2343750	Cakebread Sauvignon Blanc	FL	5306.67
532	11602750	Cloudy Bay Sauvignon Blanc	TX	7442.46
534	11602750	Cloudy Bay Sauvignon Blanc	TX	7442.46
915	2343750	Cakebread Sauvignon Blanc	FL	5306.67
1504	4350750	Simi Sauvignon Blanc	CT	5487.00

Total Wine & MORE

Opportunities

To improve the overall quality of predictions and to improve the assortment recommendations, there are various steps that we can take in the future -

- Access alternate and more in depth data about external sales. This will allow us to look at the bigger picture in terms of the entire market.
- Collect and include in the modelling, data beyond sales data. Include product characteristics like Manufacturer, Alcohol %, Flavor Profile, etc.
- Use sales data across years to build a more holistic model that will account for any price/sales variations
- Include Social Media Sentiment Analysis for products to identify trends