Report: French Wine Company ML Project

Chrys Ngouma

12. februar 2024

Abstract

This project offers a comprehensive analysis of the US wine industry for a French wine company. Utilizing a data-driven approach, our goal was to extract actionable insights critical for the client's market implantation strategy. Focused on the broader US and Californian markets, our analysis encompassed segmentation analysis, gross margins, sales volumes, consumption trends, and predictive forecasting. While successful in multiple facets, we faced challenges in forecasting gross margins due to data limitations. Our findings shed light on market trends, offering valuable insights for informed decision-making and strategy formulation.

1. Introduction

This project extensively explores the US wine industry on behalf of a French wine company. Employing a robust data-driven methodology, our primary aim is to extract actionable insights pertinent to both the Californian and broader US markets. These insights are crucial for empowering our client with a comprehensive understanding of the market landscape, aiding the identification of opportunities.

Central to our client's objectives lies a need for segmentation analysis—an essential component guiding their pricing strategy. Additionally, our client seeks comprehensive insights into gross margins, anticipated sales volumes, consumption trends, and predictive forecasting. These facets form the foundation of our analysis, focusing not only on comprehension but also anticipation of US wine market trends.

This report highlights key insights derived from our data analysis. Furthermore, it lays out the methodology employed in constructing and evaluating three distinct forecasting models, tailored to address the client's needs. However, it is essential to note that while our experiments yielded promising results in forecasting expected sales volumes and consumption trends, limitations arose in modeling gross margins, elaborated further in this report.

2. Approach

In this section, we describe the data-driven approach taken to solve the client's problem. We follow a data science methodology as described in Figure 1.

1. **Business Task Definition:** The initial step involves a comprehensive understanding of the client's needs and challenges, outlining subtasks and how Machine Learning (ML) application could address the core problem.
2. **Data Preparation:** This phase involves the gathering, cleaning, and processing of the data, handling missing data and inconsistencies, alongside conducting Exploratory Data Analysis (EDA) to grasp data distribution and correlations.
3. **Model Planning:** Here, the focus revolves around decisions on model selection appropriate to solve the client's problem. This stage includes metrics selection while considering a potential need for extra data.
4. **Model Building:** Implementation of the selected models to address the identified subtasks.
5. **Model Evaluation:** This step consists of an evaluation of the models built using the different metrics selected, as well as an optimization and finetuning the hyperparameters.
6. **Insights Formulation:** Interpreting the results of all experiments as well as the findings from the EDA.
7. **Report:** Finally, the last stage consists of summarizing all completed work, insights drawn, and experimental outcomes in a report.

3. Datasets

In this section, we describe the data used for our analysis and development of the selected models.

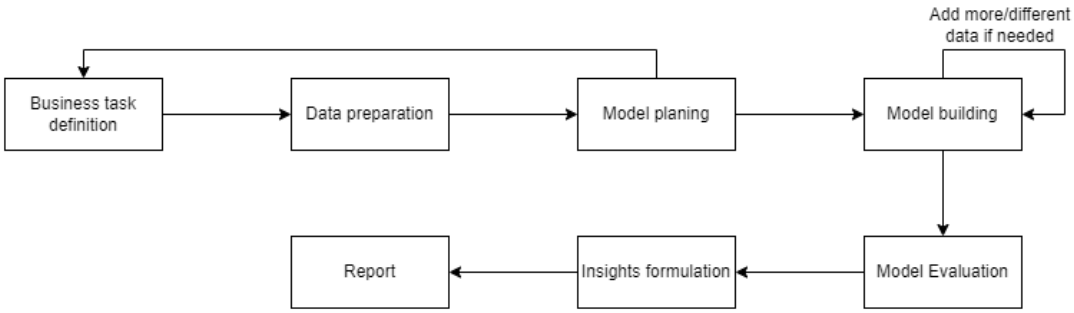


Figure 1: Flowchart of the approach taken to solve the client's problem.

3.1. Marketing Campaign

We use a dataset containing data about a marketing campaign. The data was published in 2019 on Kaggle and combines both demographic information about customers as well as how much they have spent on certain goods over 2 years (including wine products). More specifically, we are interested in the information about the customer year of birth, level of education, income, household composition, marital status, as well as how much they spent on wine. In total, the dataset has 2240 entries.

3.2. Wine Market Statistics

We also combined datasetss, mainly from statistica.com, which included the following time series:

- Gross margin of premium wineries in the US (2002-2022)
- Market share of the US alcohol industry by beverage (2000-2022)
- Total wine consumption of the US (2005-2022). We augmented this time series dataset by adding more historical data from wineinstitute.org (1934-2022).
- US and Californian wine production (2006-2022)
- Sales volume of wine in the US (1999-2021)
- Californian wine market value in the US (2006-2022)

The data of the statistics we want to forecast ranges from 2002 to 2022, which is only 20 datapoints. It is therefore important to select forecasting approaches that work well with small datasets.

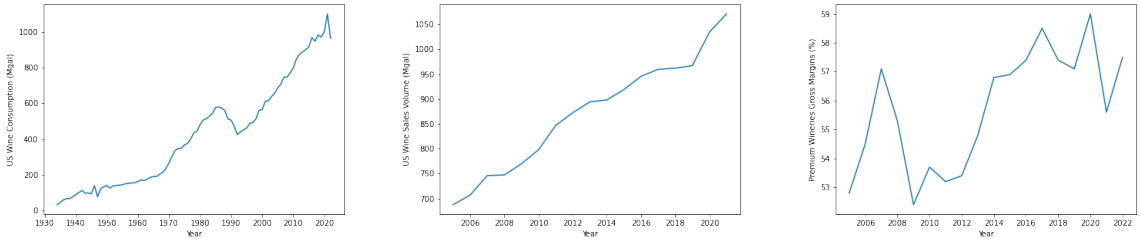
4. Predictive Models

This section presents our exploration of different methods with potential applications in solving the different facets of client's problem. The models selected are aimed at both market segmentation and forecasting, each tailored to address the client's needs and requirements.

4.1. Market Segmentation

The main objective of market segmentation revolves around uncovering cohesive clusters withing customer spending behaviors. Our aim is to group customers with similarities in their purchasing patterns and uncovers what they have in common in terms of demographics. To achieve this, we explored a few clustering algoritms such as KMeans [4], MeanShift [1], and DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [2].

1. **KMeans:** This method serves as our baseline. KMeans is a simple clustering algorithm aiming to partition data into distinct clusters. It is scalable and works well with spherical clusters. However, it is limited as it requires to specify the number of clusters beforehand, it is sensitive to the initial placement of centroids, and struggles with nonlinear or irregularly shaped clusters.
2. **DBSCAN:** identifies clusters based on density, grouping points into clusters while marking outliers as noise. It is robust against outliers and can detect arbitrary shaped clusters. In addition, this method can identify clusters of varying shape and sizes and dies not require to specify the number of clusters beforehand. On the downside, DBSCAN is very sensitive to its hyperparameters `epislon` and minimum points.
3. **MeanShift:** MeanShift can detect irregularly shaped clusters and adapt to varying densities. It is robust against outliers and also doesn't require specifying the number



(a) (b) (c)

Figure 2: (a) US Wine Consumption (b) US Wine Sales Volume (c) Premium Wineries Gross Margins

of clusters in advance. However, as limitation, this method is computationally expensive, sensitive to its hyperparameter `bandwidth`, and may converge to local optima depending on initialization.

4.2. Forecasting Models

Our forecasting models aim at predicting essential metrics such as gross margins, sales volume, and expected consumption trends within the US wine market. For this project, we focused on various statistical methods tailored to forecast each metric individually. Some of the techniques explored included Exponential Smoothing [3], Autoregressive Integrated Moving Average (ARIMA), and Dynamic Linear Models (DLM). These methods are uni-variate so we did not use any additional feature variables for most of them. Additionally, a simplistic Linear Regression model using the 'Year' as feature to capture temporal trend within each variable, was used as baseline model.

1. **Exponential Smoothing:** This method captures and forecasts time series data with both trend and seasonal components. It adapts to changing patterns over time and is computationally efficient. However, it requires consistent and stable historical data to maintain accurate forecasts, struggles with irregular or complex patterns, and is sensitive to outliers.
2. **ARIMA:** This is a powerful time series forecasting model that includes autoregression, differencing, and moving averages. It accommodates various time series structures, including trends, seasonality, and non-stationary data, and suitable for complex patterns. On the other end, ARIMA is limited by its requirement for carefully selecting hyperparameters `p`, `d`, and `q`, it struggles with small or limited data, and is not ideal for irregular patterns.
3. **DLM:** These are Bayesian state space models used for time series analysis, integrating multiple sources of information. They are quite effective but also more difficult to implement as they require a deep understanding of model components and domain expertise for accurate modeling.

5. Experiment

5.1. Metrics

In the training and evaluation of our selected forecasting models, we considered Mean Average Error (MAE), Mean Squared Error (MSE), and R-Squared (Coefficient of Determination). Meanwhile, we use dimension reduction and data visualization to assess the quality of the clusters produced by our clustering models for market segmentation.

Both MSE and MAE are suited for forecasting. MAE measures the average absolute differences between predicted and actual values thus giving an idea of the magnitude of the errors without considering their direction. MSE measures the average of the squared differences between predicted and actual values and punishes larger errors more severely due to squaring. On the other hand, R-squared measures the proportion of the variance in the dependent variable (target) that is predictable from the independent variables (features), indicating how well the model fits the data.

5.2. Cross-Validation

Cross-validation is an effective and robust technique in machine learning, serving to gauge a model's performances. It diverges from a single train/test split by segmenting the dataset into multiple subsets or folds. Each fold is then alternatively used for training the model while the remaining ones become the validation sets. We used a split ratio of 0.2 (test set being 20% of the data) as it allowed for large enough training sets to learn patterns while still having decent sized validation sets. Performance metrics are computed for each run and averaged across the iterations to derive a comprehensive assessment of the model's generalization

capabilities and detecting issues like overfitting or underfitting.

In the case of time series, it is imperative to preserve temporal order. Data in the training sets cannot be later than the ones in the validation set. We then employ a modified but similar approach named Time Series Cross-validation—implemented in Python. In each fold, we ended up with a test set of 5 datapoints.

5.3. Results

This section serves the purpose of presenting the results of training and testing of the different model implemented.

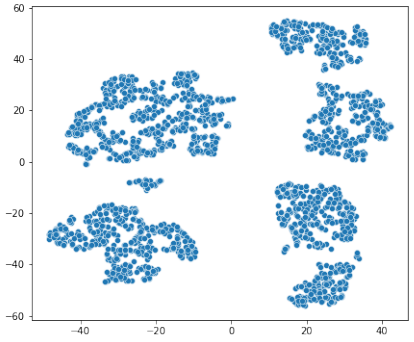
5.3.1. Market Segmentation

We performed market segmentation using the marketing campaign dataset (*cf.* Section 3.1) which contains demographic data about customers as well as how much they spent on wine. From the dataset, we selected/extracted data such as *Customer_Age*, *Education*, *Income*, *Kidhome* (number of small children), and *MntWines* (total amount spent on wine). Those were the variables having the strongest relationship with our dependent variable *MntWines* (*cf.* Section 6). Figure 3 shows a plot of the dataset after dimension reduction with t-NSE. We can visually identify 4 to 7 well defined clusters.

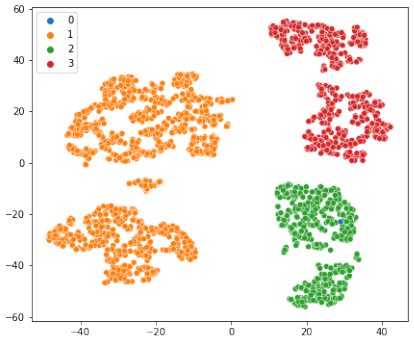
Out of the three approach considered, DBSCAN show the best results. The clusters identified by this approach fits our dataset perfectly (*cf.* Figure 4) compared to KMeans and MeanShift. An initial analysis of the clusters showed that the model used *Education* as the main criteria to group customers. Within each clusters, the customers shared nearly identical attributes otherwise. Given that *Education* doesn't have such a significant role in how much a customer spends on wine, we excluded this variable and reran the clustering algorithms. This time, a 2D plot of the data show two distinct clusters (*cf.* Figure 4). The final two clusters identified by DBSCAN are presented in Figure 4.

We summarized in Table 1 statistics related to demographics and spending patterns of each cluster. The results show that `cluster 1` are the consumers that spend the most on wine ($463.91 on average). They are characterized by having a higher income than `cluster 2` (73.6% higher), their group is composed of significantly more seniors as a percentage, and they do not have small children. On the other end, we have `cluster 2` who are characterized by a lower income, much lower percentage of seniors, and tend to have small children.

Our market segmentation analysis suggests that age, income, and presence of small children has

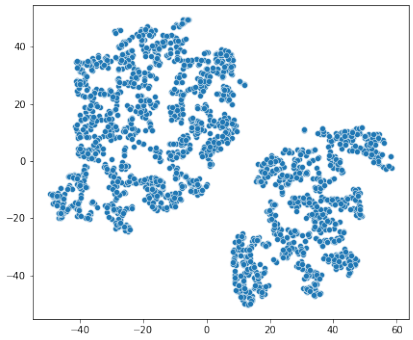


(a)

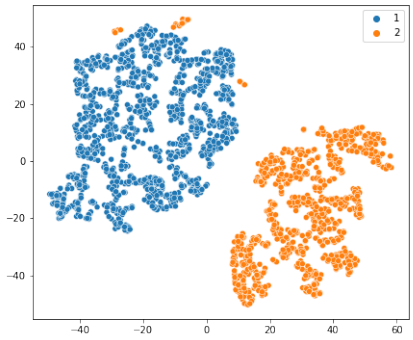


(b)

Figure 3: (a) 2D plot of `marketing_campaign` data with *Education* (b) DBSCAN clusters with *Education*



(a)



(b)

Figure 4: (a) 2D plot of `marketing_campaign` data without *Education* (b) DBSCAN clusters without *Education*

an impact on wine customers spending patterns. Meanwhile, education doesn't have a significant impact.

5.3.2. Forecasting Models

We employed multiple methods to implement forecasting models to predict wine sales volumes, expected wine consumption, and gross margins of wineries. We used the appropriate datasets from the wine market data (*cf.* Section 3.2). Figure 2 shows the trends in time series studies.

Besides ARIMAX, the methods considered used only the independent (target) variables to make predictions. In the case of ARIMAX, we added dependent variables (features) to the model. Our features included the variables having the strongest relationship with the targets (*cf.* Section 6). Tables 2, and 3 summarize the performances our forecasting models predicting sales volumes, and consumption respectively. Overall, Holt-Winters Exponential Smoothing shows the best performances when predicting the sales volumes. MAE is relatively and the model explains 87% of the data. Similarly, our forecasting model to predict wine consumption performed well but this time with ARIMA which explains 77% of the data.

In terms of hyperparameters, although our data is not seasonal, a `seasonal_periods` of half the train size gave the best results after cross-validation of Holt-Winters for the Sales Volume forecasting model. Meanwhile, we use `p=1` (number of lag observations), `d=2` (number of times the raw observations are differenced to make the time series stationary), and `q=1` (order of the Moving Average (MA) part of the model) as parameters of the ARIMA model forecasting the expected wine consumption.

On the other end, none of the models tested gave a high enough R-squared score but rather negative values. This means that the model performs worst than an horizontal line. Considering the values for gross margins in our dataset fluctuate between 50 and 58, MAE of 2.05 is not convincing either. This could be due to the very limited amount of data available. We elaborate more on this in Section 7. Table 4 summarizes the performances.

6. Discoveries and Insights

In this section, we summarize the statistics and findings from our analysis of the various datasets studied for this project.

6.1. Demographic Data

Our data analysis of the marketing campaign dataset revealed how certain demographic factors influence

Table 1: Summary statistics of clusters

Clusters	Cluster 1	Cluster 2
Wine expenses (mean)	463.91	105.64
Wine expenses (median)	410	27
Income (mean)	62264.90	38870.68
Income (median)	64509	37150
Num Adults	860 (68.74%)	781 (87.45%)
Num Seniors	391 (31.25%)	112 (12.54%)
Has Small Children	0 (0%)	893 (99.89%)
Undergraduate	643 (51.4%)	467 (52.3%)
Master's degree	311 (24.86%)	247 (27.66%)
PhD	297 (23.74%)	179 (20.04%)

Table 2: Summary of the performances of the Sales Volume forecasting model

Model	MAE	MSE	R^2
Baseline	16.51	469.069	0.69
Holt-Winters	**11.16**	**172.02**	**0.87**
ARIMA	15.93	445.25	0.61
DLM	17.46	494.07	0.69

Table 3: Summary of the performances of the Expected Consumption forecasting model

Model	MAE	MSE	R^2
Baseline	24.58	805.43	0.68
Holt-Winters	23.65	726.82	0.71
ARIMA	**20.91**	**590.34**	**0.77**
DLM	24.23	774.67	0.69

Table 4: Summary of the performances of the Gross Margins forecasting model

Model	MAE	MSE	R^2
Baseline	2.88	14.32	-14.03
Holt-Winters	4.31	24.24	-22.65
ARIMA	2.43	9.21	-7.78
DLM	3.52	20.5	-20.49
ARIMAX	**2.05**	**5.69**	**-4.11**

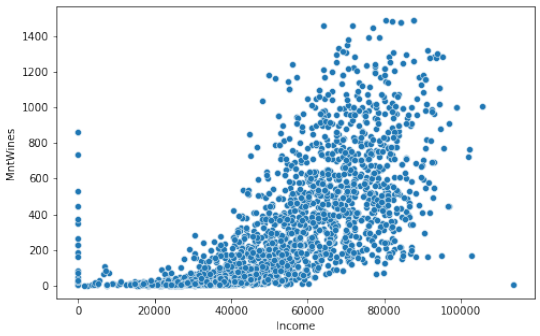


Figure 5: Relationship between Income and amount spent on wine (*MntWines*). There is a strong correlation between the two variables (Spearman coefficient: 0.82).

spending patterns on wine products. Table 5 summarizes statistics about the amount spent on wine per demographic groups. On average, certain demographic groups tend to spend more than others. For instance, Seniors customers have the tendency to spend more than young adults and adults. This could be due to age or a generational factor. It could be that older generation are more drawn to wine while younger generation favor spirits and beer. Overall, age group (senior vs adults), education level, and presence of small children in the household signal the most significant difference in spending patterns. Customers who are Seniors, have PhD, or no small children tend to spend more than the other groups.

In addition to these demographic groups, income also has a significant influence on how much customers spend on wine. Figure 5 shows that the higher the income, the higher customers tend to spend on wine. This trend is expected as people with more financial capability can spend more on anything.

Table 5: Average amount spent on wine per demographic groups. Here *Adults* include customers 21-59.

Demographic group	Avg amount spent	Median amount spent
Adults	291.40	158.0
Seniors	384.89	303.0
Undergraduate	285.78	185.0
Master's degree	288.70	155.0
PhD	406.79	284.0
Has small children	107.06	27
No small children	462.08	407.5
Widow	374.09	336.5
Divorced	331.92	189.0
Married	308.62	183.0
Together	315.68	199.0
Single	300.30	163.0

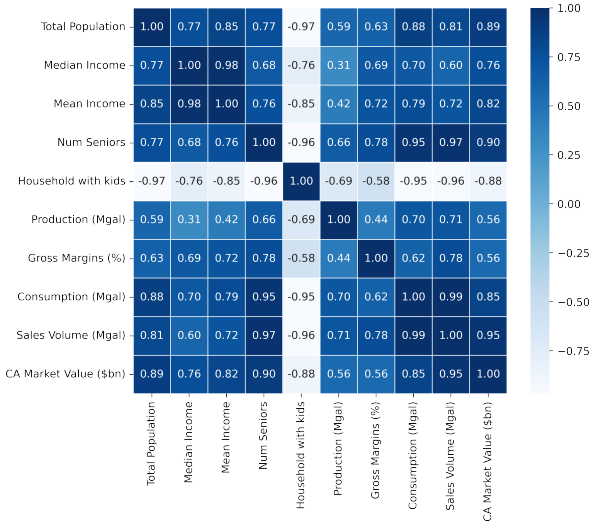


Figure 6: Correlation table between variables.

6.2. Wine Market Statistics

In this section, we present the insights and observations made during the analysis of the wine market dataset. We augmented this dataset with US Census demographic data[1]. Overall, our target variables, wine sales volume, expected consumption trends, and gross margins of premium wineries, increase over time. This could be due to multiple factors, including general population growth. Figure 6 shows the correlation table between our studied variables and other statistics including demographics and market statistics.

We find that certain variables such as *total population* have a strong correlation with each other the three target variables. This is expected—as the population grows, the demand for most products also increases. Similarly, the higher the income of customers, the more they can afford to spend on any given product, including wine. On the other end, we observe a strong inverse correlation between our target variables and the number of household having small children. This observation is supported by the findings in Section 6.1.

7. Challenges

Along this project diving into the US wine market, we faced some challenges. The main challenge was data scarcity, in particular data about gross margins of US wineries. Although the dataset available covers 20 years from 2002 to 2022, this data is annualized hence only has 20 datapoints. This amount of data pose a challenging when training and, especially, testing our forecasting models. As we adopted cross-validation as a robust technique of measuring our models performances, 20 datapoints proved to be very limited. This could explain the poor performances of our forecasting models predicting gross margins. The model wouldn't have much to learn from.

8. Conclusion

Our exploration of the US wine industry for our client has yielded significant successes and noteworthy challenges. The segmentation analysis revealed interesting patterns in customer spending behaviors, emphasizing the influence of age, income, and household composition. Furthermore, our forecasting models exhibited promising performances in predicting sales volumes and consumption trends but faced limitations in effectively forecasting gross margins due to limited data.

Despite encountering limitations, our analyses provide our client with actionable insights, enabling informed decision-making in navigating the dynamic US wine market. Addressing data scarcity remains

[1] www.census.gov/

critical for refining forecasting models and ensuring robust outcomes.

References

[1]Y. Cheng, ≪Mean shift, mode seeking, and clustering≫, IEEE transactions on pattern analysis and machine intelligence **17**, 790–799 (1995).

[2]M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al., ≪A density-based algorithm for discovering clusters in large spatial databases with noise≫, in Kdd, Vol. 96, 34 (1996), pp. 226–231.

[3]E. S. Gardner Jr, ≪Exponential smoothing: the state of the art≫, Journal of forecasting **4**, 1–28 (1985).

[4]J. A. Hartigan and M. A. Wong, ≪Algorithm as 136: a k-means clustering algorithm≫, Journal of the royal statistical society. series c (applied statistics) **28**, 100–108 (1979).