**Solving technical problems**

**Data Analysis and Visualization**

**{Title}**

Applying the CRISP-DM Standard to a Real-World Problem – The Winemaking Use-Case

**{Description}**

A step-by-step guide through the process of solving a real-world problem by using the CRISP-DM standard.

**{Objectives}**

* Learn to apply the CRISP-DM methodology to solve a real-world problem
* Learn to make difficult and complex decisions at each CRISP-DM phase, or even sometimes revisit previous phases

**{Contents}**

 **Source: Generated by ChatGPT using DALL-E**

**Introduction**

When approaching a business problem where the solution relies upon understanding and making use of data available to the business, having a framework to allow common terminology and effective communication throughout the project lifecycle is useful.

A popular framework for this that has been around since the late nineties is called CRoss Industry Standard Process for Data Mining (CRISP-DM) and breaks the process into six stages or phases. This framework is often modified to fit a data team’s approach, with more weight on areas of importance or broken down into custom substages with artifacts relating to a team’s project management and software development tools.

While these stages seem intuitive when reading through them, depending on the problem in question, there may be some difficult and complex decisions at each stage, or even the need to revisit previous stages, like any iteration-based approach. For this example, we will use a fictional Winery business and its desire to crush its competition. We source the data from Kaggle’s red and white wine datasets.

**Phase 1: Business Understanding**

***Business context***

Grape Expectations Winery has acquired data from a trusted source that contains the chemical composition of wines and their quality rating. They want to use it to increase their revenue and reputation as the greatest wine purveyors in the business.

One business initiative they have is to win the local Timboon Region Wine Festival, in which they have been coming in consistent 2nd and 3rd place. They believe with some data mining and analysis they can have an edge over the competition.

The success criterion is winning the competition. To do that they need to do two things, know how to produce a top-quality wine, and also know which wine to put forward for the competition, as they always produce multiple wines and select one for entry.

***System requirements***

System requirements are not complex as this is a standalone modelling project for now which has a low enough cadence and data volume to be performed on their local systems by some educated employees. This does mean the model and code should be easy to use by a slightly technical person if it is to be deployed, however, for this engagement the data expert will be onboard for the entire lifecycle. If deployment is required that will be a separate business case.

**Phase 2: Data Understanding**

***A look at the data***

A screenshot of a computer

Description automatically generated

**Sample of the wine quality dataset**

The data available is a CSV with a series of quantitative measurements of a particular wine and a rating of its quality by a wine evaluation expert. This brings us to an important assumption to validate with the business.

We assume the judges in the competition this year will have a similar assessment of quality as the judges who produced these quality scores.

If this is not the case, then our data may be considered useless as there is no dependency between the target score in the sample data and the target score we are trying to predict.

***Data profiling***

A simple way to start exploratory data analysis (EDA) is by profiling each column. In this case, we use Python to automate this process and generate a visualization that can serve as a helpful reference to validate our understanding of what each column is and also provide further insights about the domain, guiding us in the next steps.

***Correlation***

A good start is a correlation plot of variables. Let’s see what contributes highly to quality.

A chart of different colored squares

Description automatically generated

**Heatmap of variable correlation, dark blue indicates positive correlation and red indicates negative correlation. We see alcohol, sulfates and volatile acidity have the highest correlation with quality.**

***What to focus on***

Before looking at the correlations and any other results of the profiling, we need to consider what is worth focusing on. To do this, consider what variables are rigid and what the winery has some control over. For example, we may see a high correlation between pH and quality, but maybe pH is highly dependent on the altitude the grapes are grown at, which is difficult to change, to say the least. We would then incorporate this into how we approach the problem, maybe by finding the pH range available to the winery and then filtering the dataset to only wines that fit into this range.

This highlights the importance of linking observations of the data to relevant business processes and objects. The goal of most data mining projects is to influence a decision that is tangible in the real world, which means the data must relate to the factors in that decision.

For instance, the wine density may turn out to be something the winemakers have a lot of control over, looking further into this, the density is highly correlated with the alcohol content and several other factors that makes density itself not the variable to focus on changing but the contributing factors.

***What are we trying to predict?***

Assessing the target column indicates that most wines fall in the 5-6 quality, which means we don’t have a lot of information for a model to confidently know what makes a wine rank 8/10 in quality.

A blue graph with white text

Description automatically generated

**Histogram of wine dataset row count by quality**

A few factors can influence the model and algorithm selection, business context, system requirements, and performance. All of this is important to consider when in the data exploration phase. This is likely a phase that will be revisited many times throughout the lifecycle. We have only really scratched the surface of what can be explored and analyzed in this stage. Often the data exploration phase takes longer than expected.

***Outcomes***

At this stage one might suggest to the vineyard that getting a model that predicts 8/10 wines accurately may not be possible. So, to simplify the use case and get the most bang for your buck, you could convert this into a problem of predicting a 7+ quality, and then at least the number of wines for selection to go to the competition is lowered and the chance of them being of good quality is higher.

**Phase 3: Data Preparation**

The data preparation we are doing will have two stages: common data preparation and model-specific data preparation.

***Common data preparation***

Common data preparation will focus on generic data cleaning, such as dealing with nulls, normalization, and feature engineering. We can create some interesting features with this wine dataset, after a quick online search appears that the balance between acidity and sugars is something that impacts the quality of wine, we can create a new feature that is equal to the acidity divided by the sugars. After we do this, let’s check our correlation analysis again to see if any features have a high correlation.

A few slightly correlated features we have engineered may improve our model’s output. We also see total\_acid and acid/density showing the same characteristics, this could be due to density being ~=1, and normalizing the features could be a way of improving this.

A diagram of different types of substances

Description automatically generated

**Heatmap of engineered feature correlation, the dark blue indicates positive correlation and red indicates negative correlation.**

***Model specific preparation***

Model-specific preparation includes any encoding required for the classification algorithm we are using, one-hot-encoding for example. Or transforming the target column, into a binary 0 and 1 where 1 represents 7 or greater quality. In this project we have a single, static and relatively clean dataset, so we do not need to do much preparation, however, in general data preparation is one of the most time-consuming stages. Particularly if you haven’t been thorough with the data exploration.

**Phase 4: Modelling**

***What type of problem is this?***

We focus here more on the CRISP-DM lifecycle and all the things one can consider along the way, so diving into what algorithm would actually be best is not our technical focus now. As we saw in the data exploration stage, we are trying to predict an integer from 1-10 where most values are 5 or 6. We will try a few techniques for this and see how it performs. We will also take into account our findings of there being very few high-quality wines to train from and see if we can get a better model by reducing the quality value to a 1 or 0 with 1 being 7 or higher quality.

***Linear regression***

This is similar to the classic Excel trendline that you may be familiar with, except we need to round the prediction to the nearest whole number before assessing the results. Plus, we are fitting the data to many variables rather than just the one that is common in the Excel use. To assess the results we will use a confusion matrix as it is an easy way to see what the model predicts and what the true values of quality are. The linear regression results are below:

A chart with numbers and a number in a square

Description automatically generated with medium confidence

**Linear regression confusion matrix, predicted vs true.**

Let’s check out the results after resampling:

A chart with numbers and a number in a square

Description automatically generated with medium confidence

**Linear regression confusion matrix, predicted vs true after resampling.**

We see this helps pull some of the 7 values up into their correct position. It also increases how many 5s and 6s get incorrectly valued higher.

***Multi class***

This is an algorithm that treats each quality value as an independent label and tries to train a model to predict them correctly.

A chart with numbers and a number in a square

Description automatically generated with medium confidence

**Multi class model confusion matrix, predicted vs true.**

***Ordinal regression***

This technique makes use of the ordering information in the class attribute, and it has been shown that this often improves accuracy compared to treating the class values as a simple, unordered set.

A chart of a number and a number

Description automatically generated with medium confidence

**Ordinal regression model confusion matrix, predicted vs true.**

***Binary classification***

As mentioned, we can convert the target to just predict whether a wine is 7 or above in quality. In doing this we now have all the binary classification options available to us, in this example, we use a model called XGboost and optimised it to this dataset.

A chart with numbers and a few colored squares

Description automatically generated with medium confidence

**XGboost model confusion matrix, predicted vs true.**

**Phase 5: Evaluation**

***A holistic and retrospective analysis***

While in the modeling selection, we try and get the model that produces the best output according to standard model evaluation techniques, this evaluation stage is more focused on the business perspective. For our case, we consider all the assumptions we have made, then consider the output of the model and whether it is suitable for influencing a business decision or providing value.

Good quality data is the key to making good business decisions. The switch to a binary prediction of good quality to reduce the number of wines to select from is valuable, plus the assessment of the important features such as alcohol content, sulfates and volatile acidity.

A good recommendation would be to suggest the binary classification model as it has a high recall. This means of all the 7+ wines, it correctly identifies the most, with a slightly higher false positive rate. If we assume there is a step after the prediction results are obtained to have experts within the winery assess the identified high-quality wine, then this gives us the best chance to not miss the best candidates.

**Phase 6: Deployment**

***Low complexity in this Case***

The deployment for this could be as simple as giving the chosen model code/notebook to the winery and educating them on how to input the variables of the wine they want to assess and get a prediction of quality. So, they can self-serve the analysis. However, some Machine Learning Operations (MLOps) best practices around cataloging training data, saving model artifacts and some versioning would save time in the future.

If they wanted to retrain the model on new data or edit its functionality in any way, this would require some education on how this could be done or a lightweight repetition of the CRISP-DM lifecycle with any new requirements in mind.

This project did not require a complex deployment step, though if it did, we would engage an MLOps lifecycle approach to deploy a model that can be easily, validated, maintained, retrained and monitored. This could be a topic for a future post.

**Final Thoughts**

We have gone through a lightweight example of using an industry-standard framework for producing a model that can generate business value through identifying high-quality wines to be the focus for further assessment. However, we only scratched the surface of statistical analysis, modeling and business analysis of the problem statement. This article aims to highlight the benefit of breaking down a data project into defined stages and using common terminology, yet also revealing the gaps that a generic framework may not prescribe artifacts or tasks for, these gaps, therefore, need to be filled by people with subject matter expertise and ability to ensure project success.

We also didn’t try some popular classification techniques such as Neural Networks, Naive Bayes or a Support Vector Machine.

Another pertinent point in the design and deployment of a ML system is the importance of being thorough in the early CRISP-DM lifecycle stages, as questions that arise when designing a production system may not have been fully answered when doing the proof of concept that only focused on the design of the model itself.

For instance, when defining a set of rules for model input validation, it is useful to know what assumptions on the data are made for the model to generate relevant and useful outputs. For example, in the wine quality, we did not explore what constraints were on some of the columns, like the pH column. The model may equate a high pH with a low level of sulfates with high quality, and if the pH comes in at 7, then it may indicate a high-quality wine, when in fact a pH of 7 is not within the expected range and the max in the training data is 4. This would be clarified by a domain expert as well, pH of 7 means completely neutral acidity which doesn’t make sense for an alcoholic substance.

In conclusion, breaking down a data project into defined stages and employing common terminology can bring numerous benefits to its execution. It facilitates clear communication, promotes collaboration and provides a structured approach. However, it’s important to acknowledge that individual subject matter expertise and validating data assumptions and constraints with domain experts are key to reaping the full benefits of data-driven initiatives.

**{Activities}**

1. Analyze the described application of the CRISP-DM standard to the winemaking use case and try to identify if there is room for improvement!
2. Propose some possible improvements to the described data mining analysis scenario involving winemaking!

**{Questions for self-assessment}**

1. What is the primary objective of Grape Expectations Winery in using CRISP-DM?
   * To win the local Timboon Region Wine Festival.
   * To increase overall wine production.
   * To reduce the cost of wine production.
   * To expand their product range.
2. What type of data does Grape Expectations Winery analyze in CRISP-DM's Data Understanding phase?
   * Chemical composition of wines and their quality rating.
   * Demographic data of wine consumers.
   * Global wine market trends.
   * Historical weather data affecting vineyards.
3. In the Business Understanding phase, what is the success criterion for Grape Expectations Winery?
   * Winning the local wine competition.
   * Achieving a specific revenue target.
   * Receiving a high rating from wine critics.
   * Increasing market share in the region.
4. What assumption is validated by the business in the Data Understanding phase?
   * The competition judges will have a similar assessment of quality as the data judges.
   * The wine's quality improves with age.
   * The local market prefers red wine over white wine.
   * The wine production process can be significantly automated.
5. What approach is used in the Data Understanding phase to start exploratory data analysis (EDA)?
   * Profiling each column of the dataset.
   * Conducting customer surveys.
   * Analyzing competitors' wine quality.
   * Reviewing historical sales data.

**{Questions for official assessment}**

1. Which feature did the winery decide to focus on based on correlation analysis?
   * pH level of the wine.
   * Density of the wine.
   * Altitude at which grapes are grown.
   * Sugar content in the grapes.
2. In the Data Preparation phase, what new feature is created for the wine dataset?
   * The ratio of acidity to sugars.
   * Average grape size.
   * Total number of sulfates.
   * Fermentation duration.
3. What modeling technique is first tried in the Modelling phase?
   * Linear regression.
   * Decision tree analysis.
   * Neural networks.
   * Cluster analysis.
4. In the Evaluation phase, what was a key consideration for the winery?
   * The model's ability to influence business decisions or provide value.
   * The cost of implementing the model.
   * The time taken for the model to run.
   * The scalability of the model for larger datasets.
5. What approach is taken for deployment in this case study?
   * Educating winery staff on using the model code/notebook for self-serving analysis.
   * Fully automating the wine quality assessment process.
   * Outsourcing the model management to a third-party data firm.
   * Implementing a cloud-based real-time analysis system.