First Steps Of The Project

Project C7: KAGGLE-WINE-REVIEW

Team members: Henri Sellis, Juhan Tamm (Both in group 5)

Github repository link: https://github.com/Juhantam/WineProject (public repository)

Business understanding

Identifying your business goals

Our project developed from the fact that there is a large number of detailed information about wine reviews available. In addition to details about the wine and the grapes that the wine was made of, the data also contains information about the sommeliers who conducted the review, and an exact description of the wine by the sommelier. This information is the basis for our business goals.

Our first business goal is to find out details about the highest rated wines. This information is potentially very useful for companies who buy or sell wine. In addition, it helps people choose a wine that they would most likely enjoy thoroughly.

Our second business goal is to build an algorithm that could predict the sommelier's rating for the wine from the rest of the information. This algorithm could be used to automate some parts of the wine reviewing process. Using this specific dataset, the algorithm would still rely on the wine description written by professional sommeliers, but in the future, the descriptions could be replaced with different specific details about the wine's taste, which would even allow the algorithm to work completely without the sommeliers.

Assessing your situation

All the data that we have for our project is available on a platform called Kaggle. The data can be used completely free of charge as long as the use is non-commercial. The hardware that we have at our disposal is provided by the University of Tartu, more specifically HP's Elitebook laptop computers. These laptops are quite powerful, the only downside is that they have Intel's integrated graphics, but that will not affect our project almost at all, since the required processes mostly need processing power and RAM, not graphical power. The software solution that we are going to use in this project consists of multiple programs that work very nicely together. The most important of them is Python 3, this is the core of all our work. It allows us to apply machine learning and data analysis techniques to our data in a very neat fashion. The program that helps us manage all our code is the Jupyter notebook. Jupyter notebook is very useful for any kind of data science project because it

provides a way to easily run our code in different parts and is in addition very convenient for displaying various graphs. The last important tool is Github, which is used for version control and sharing work. In addition to all these technical features at our disposal, we can also seek guidance from our instructors at the University of Tartu.

The project has to be finished by the 17th of December 2020. The only legal constraint is that the data cannot be used commercially and security-wise there are no constraints at all. The finished work is acceptable if it meets our business goals.

There are not too many contingencies when it comes to our project. The worst that could happen is losing electricity, which would make it very difficult to continue working. Another hurdle would be the loss of internet connection because that would significantly hinder our work speed, but at least we could still keep working.

There are no specific business nor data-mining terms in our project.

Our project will not cost us anything, with the only exception of time consumption. However we will also not make any materialistic profits from this project, instead, we will have a lot of valuable information in the end.

Defining your data-mining goals

Essentially the only deliverable in our project is the final poster presentation in December.

Firstly, we are hoping to create a model that could accurately predict the wine ratings from the wine description. Secondly, we are hoping to gather a detailed overview of the best wines and their features.

We will consider our data-mining goals fulfilled when at least one of our created models has an accuracy of at least 80%.

Data understanding

Gathering data

To reach our project goals, we need data that would describe the wine reviews, wines, and sommeliers in great detail. The data format that would suit us the best is '.csv'. These requirements are fulfilled by our data.

We have confirmed that our data exists and is completely available to us. We have created a Python Notebook file, loaded the data in, and confirmed that the data indeed has all the qualities and features that we expected. It is just as usable as any other data from our previous homeworks.

As for selection criteria, we will be using almost everything in our dataset. We have only one dataset with 12 features, and we can find a use for every feature included there. The only feature that we most

likely will not use is “taster_twitter_handle” because we can already identify different sommeliers from their names and our project has nothing to do with Twitter.

Describing data

Like we mentioned earlier, the source of our dataset is Kaggle and it is in a CSV file format. In total, there are 13 columns and 130 thousand rows in this dataset. The labels of the columns are the country of origin, the description given to the wine by a sommelier, the designated vineyard where the grapes come from, points assigned to the wine by a sommelier, the price of the wine, the province or state the wine is from, the general region within the province or state, a more specific region in the general region, the name of the sommelier who rated the wine, the title of the wine review, the variety of grapes used in the making of the wine and finally the winery that produced the wine. “Country”, “province”, “designation” and “variety” will mostly be used in the statistical analysis of the wines. “Points” will be the most important feature for the machine learning part, because that is what our algorithm will be trying to predict. “Price” will also be useful in the statistical analysis, for example when we find out the overall wines with the best qualities, we can also compare their prices and find out the cheapest ones among them. “Taster_name” can be used to compare how highly certain wines are rated by different sommeliers - are every sommelier’s ratings relatively similar, or are there big differences, too?

Exploring data

“Country” is a string feature with 43 different values: the top 3 countries that are considerably more frequent than others are 1. US, 2. France, 3. Italy.
“Description” is a string feature that represents the description of the wine. Each description is unique.
“Designation” is a string feature with 37979 different values.
“Points” is an integer feature that ranges from 88 to 100. Lower point values seem to be more frequent than higher point values: 17207 reviews have “points” value 88 and only 19 reviews have 100 points, with the other point values in between.
“Price” is a floating-point arithmetic feature with 390 different values. The lowest price is 4.0 and the highest 3300.0.
“Province” is a string feature with 425 different values.
“Region_1” is a string feature with 1229 different values.
“Region_2” is a string feature with 17 different values.
“Taster_name” is a string feature with 19 different values.
“Title” is a string feature with 118840 different values.
“Variety” is a string feature with 707 different values.

Verifying data quality

We have not found any major quality issues in the dataset we are using. The only minor quality issues include a few NaN values in varying columns of the data. However, this will not become a problem in the long run, since we still have plenty of data to work with.

Planning your project

List of tasks:

1. Cleaning and preparing the data. (2 hours per person)
2. Looking for interesting correlations and statistics about the data. (6 hours per person)
3. Creating various charts, plots, and graphs about the most interesting facts and insights. (6 hours per person)
4. Training and testing various models to predict the points assigned to the wines by the sommeliers. (6 hours per person)
5. Training and testing various models to predict the price of the wines. (4 hours per person)

List of methods and tools:

Most important Python libraries for our project:

- pandas
- scikit
- numpy
- seaborn
- matplotlib

We will be using machine learning and data analysis techniques that we have learned from the lectures and homeworks. For example, when developing our machine learning models, we will first split the data into training and testing datasets to avoid overfitting.