Available online at www.sciencedirect.com

ScienceDirect

Procedia Economics and Finance 33 (2015) 256 – 268

Procedia Economics and Finance

www.elsevier.com/locate/procedia

7th International Conference, The Economies of Balkan and Eastern Europe Countries in the changed world, EBEEC 2015, May 8-10, 2015

A Statistical Analysis of Big Web Market Data Structure Using a Big Dataset of Wines.

Athanasiadis Ioannis[a]*, Ioannides Dimitrios[b]

[a]*University of Macedonia, Thessaloniki, 546 36, Greece*
[b]*University of Macedonia, Thessaloniki, 546 36, Greece*

Abstract

The web market structure nowadays it's facing the issue of "big-data". It is well known that big companies in the web like Amazon , Google , are trying to get information from their data. However similar interests have also other smaller companies. In this paper we considered a large dataset from a big company of wine, with white, rose and red wines (from Greece). Two regression techniques were considered the multivariate linear regression and logistic regression. The logistic regression gives stronger results in terms of the interpretation, outperforming the linear regression and we are expecting that others methods as neural networks and support vector machine improve the last one. We are dealing about this issue in another paper. Such models are useful to oenologist wine tasting evaluation and improve wine production. Furthermore, similar techniques can help in target marketing by modeling tastes from other markets.

Peer-review under responsibility of Department of Accountancy and Finance, Eastern Macedonia and Thrace Institute of Technology

Keywords: e-commerce; big data; R language; statistical models;linear regression; logistic regression; wines

* Corresponding author. Tel.: +0030-2310-891785
E-mail address: athang@uom.edu.gr

Peer-review under responsibility of Department of Accountancy and Finance, Eastern Macedonia and Thrace Institute of Technology
doi:10.1016/S2212-5671(15)01710-4

1. Introduction

The emerging phenomenon of "big data" nowadays increases with the big web market data and consequently affects any company that want to maintain engaged in this area, looking and searching methods to achieve useful information from its big datasets.(Athanasiadis I and Ioannides D., 2014)

As we know big companies in the web like Amazon, Google or Facebook etc. drive the era of "getting knowledge from the data". Their immense flow of data shows that with an appropriate statistical analysis and predictive analytics among these data it is possible to produce strategic business decisions.

This progress of "data analysis" produced many important tools that today even a small company can use with safety and investigate the results for its benefit. The R language it is such important tool that produce with many statistical models very interesting results, especially in the predictive area.

Such a wine company in Northern Greece producing wines and having a big market area not only in Greece but also in Europe, gave us a big dataset of its wines produced, in order to make statistical research using various models and predictive analytics. This dataset includes more than 5600 records of wines with almost all the physicochemical properties, including three types of wines (white, red and roze) and covering about 9 years of production (from 2004 to 2013).

The history of wine in Greece comes from the ancient times and it is considered in many countries around the world as a top class wine. There are about 300 varieties and three main species or types: white, red and roze. (Galanos V., 2013 [http://antikleidi.com/2013/03/23/wine-3/]) The importance of the wine in economy is enormous given that as producer is 12^{th} in the world ranking! (Karlsson B., 2012, [http://www.krasiagr.com/?p=52263]), (International Organization of Vine and Wine, 2013 Report, [http://www.oiv.int]).

The exports of the Greek wines are always through the first exporting products in Greece and among the first 10 countries in the world, varying from 600.000 to 270.000 HL and 79 to 52 millions in Euros/year through the years 2000 – 2008. (Pan-Hellenic Union of Graduates in Oenology, PANEPO, [http://www.panepo.gr/Statistics.htm])

The countries that import Greek wines are almost though the biggest markets worldwide: Germany, Usa, France, Canada, Great Britain, Belgium, Italy, China, Russia etc. (Koumakis L., 2012 [http://www.analyst.gr/2013/06/25/828/]), (Startup Greece , 2011, [http://goo.gl/6zC4lS]).

From the above indices it is obvious that a research for the wine quality in Greece it is crucial for big business decision and economic consequences. The wine quality needs to be classified and certified. Within this context wine qualification and certification prevents the illegal adulteration of the wine and assures quality for the wine market. (Cortez, 2009)

Physicochemical and sensory test are requirements for such a certified wine, with the first given by certain chemical properties indexes and mainly by the human testing the latter. Taste is the least understood of the human species and the relationships between the physicochemical results and sensory analysis are difficult and still not fully understood.(A. Legin, et.al., 2003)(D.Smith, 2006)

Often highly complex and big in volume datasets are difficult to collect, store and process. Advances in information technology and especially the statistical language R (open source and free with many libraries available) have made possible to resolve these tasks. (E.Turban, et.al, 2007)

Data mining (DM) techniques aim at extracting high-level knowledge from raw data. There are several DM algorithms, each one with its own advantages.(I.H.Witten et.al, 2005) When modeling continuous data, the linear/multiple regression (MR) is the classic approach but when we have a scaling quality variable that is our final target to predict is very useful to try the logistic regression model dividing accordingly the quality to "bad" and "good" wines.

Variable and model selections in our big dataset are critical issues for applying these DM methods. Variable selection (I.Guyon, 2003) is useful to discard irrelevant inputs, leading to simpler models that are easier to interpret and that usually give better performances. Complex models may overfit the data, losing the capability to generalize, while a model that is too simple will present limited learning capabilities.

The use of decision support systems by the wine industry is mainly focused on the wine production phase (J.Ferrer, 2008) but in our case we have both results: the physicochemical test and the sensory test as the quality score given by the company tasting system. Therefore our study to predict the quality score based on arises an important need for the company and consequently in any producing company in the market. Finding a way in an

environment of continuously massive datasets generated by the production, to predict quality score and guide business decisions accordingly.

Looking the history of these tries, in 1991 the "Wine" dataset was donated into the UCI repository (A. Asuncion, 2007). The data contain 178 examples with measurements of 13 chemical constituents (e.g. alcohol, Mg) and the goal is to classify three cultivars from Italy. This dataset is very easy to discriminate and has been mainly used as a benchmark for new DM classifiers.

P.Cortez et. al, made predictions of vinho verde wine (from the Minho region of Portugal) taste preferences, showing its impact in this domain. In contrast with previous studies, a large dataset is considered, with a total of 4898 white and 1599 red samples and 11 chemical parameters. In that paper used the Neural Network and Support Vector Machine method to investigate the dataset. (Cortez et.al, 2009)

In our paper we have a big dataset of wines, having more than 5600 records, with 2795 red, 2315 white and 494 roze. Our parameters are 13 including date of production and kind of wine (or variety). These two parameters in this study are excluded for compatibility reasons but very useful in future studies.

The paper is organized as follows: Section 2 presents the wine data, Preliminary Graphic Analysis, Exploratory Graphic Analysis, Principal Component Analysis, DM models and variable selection approach; in Section 3, the Linear Regression Analysis and Prediction and in Section 4, the Logistic Regression Analysis and Prediction is described and the obtained results are analyzed respectively; finally, conclusions are drawn in Section 5 with the comparisons of the results.

2. Materials and methods

2.1. Wine data

This study considers about 117 varieties of wines produced by a well known wine producing company of the North Greece. Most of these wines is exported mainly in Europe and in other countries. We analyze three wine variants: the White, Red and Roze (or Rosé). The data were collected from April/2004 to October/2013 tested with an official certification of the quality department of the company. Each entry denotes a given test (analytical and sensory) and the final database was exported into a single sheet (.csv).

We are interested in using statistics to understand whether a wine data having, for instance , more alcohol or sulphur dioxide taste better. During the stage of preparing the variables of the datasets a big amount of missing values in two variables: "Fe" and "Shade", forced to eliminate them knowing that they do not have impact in the final results. Another incompatibility was with the existence of the "Date" variable. The decision was for these explicit DM methods to not include this variable because do not serve for our tests. On the other hand we keep consciously in mind to use this time series with other DM methods in future to extract interesting results about the time dependencies.

Since the wine variants are quite different the analysis will be performed separately, thus three datasets were built with 2795 red, 2315 white and 494 roze. Table 1 presents the physicochemical statistics per dataset and the quality score for each of them in a scale that ranges from 1(vary bad) to 10(excellent). Fig. 1 plots the quality histograms of the three variants. Our principal goal is to identify which of these variables have a significant effect on wine quality.

Table 1. The physicochemical data statistics per wine type

	White wine			Red wine			Roze wine		
Attribute (units)	min	max	mean	min	max	mean	min	max	mean
alcohol (% vol.)	9	13.65	11.71	9.05	14.85	12.55	9	14.70	12.25
pH	0	3.98	3.38	3.02	4.07	3.50	2.98	3.68	3.29
total acidity (g(tartaric acid)/dm^3)	0	8.93	5.01	4.01	10.13	5.51	3.71	9.38	5.56
volatile acidity (g(acetic acid)/dm^3)	0.100	1.430	0.306	0.120	1.060	0.381	0.140	0.500	0.317
sugar (g/dm^3)	0.50	36	5.14	1.80	47	6.98	1.50	40	5.94
colour intensity	0.030	3.080	0.077	0.06	36.16	6.79	0.08	5.21	0.93
free sulphur dioxide (mg/dm^3)	6	65	38.5	16	67	38.21	14	68	38.32
total sulphur dioxide (mg/dm^3)	26	248	138.7	12	234	96.62	45	172	111.9

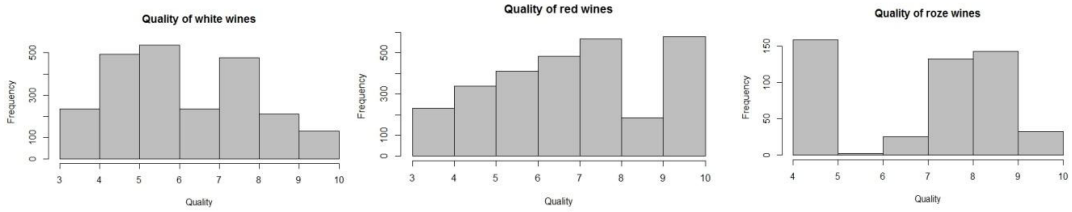

Fig.1 Quality histograms

Our principal goal is to identify which of these variables have a significant effect on wine quality.

2.2 Preliminary Graphical Analysis

Apart from the linear regression model in this study we consider also the Logistic Regression model. In this framework we divide each dataset in "bad" and "good" wines. The bad is till the quality score 6 else is a good one. In fig.2 there are the plot accordingly to this division.

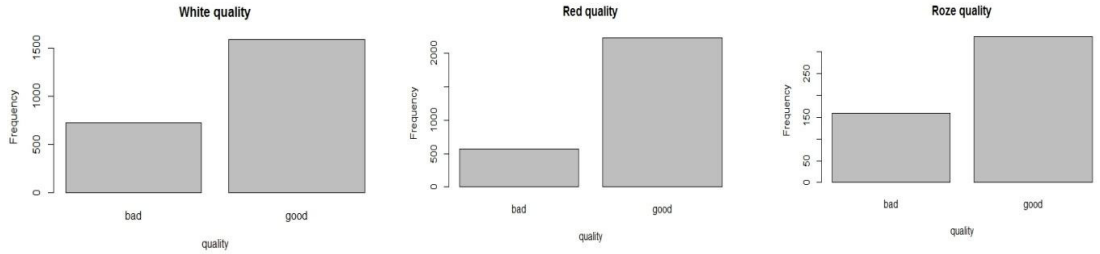

Fig. 2 Logistic division of quality

As we can see in these figures the red "good" quality is obviously near the 80% of the total, while to the other wine types are quite near 67%.

In this section our primary objective is to evaluate the effects of pH, sulphates, alcohol, sugar, and other factors on wine quality.

We are interested in identifying variables for which there is a large change between a good wine and a bad one. These variables might be a good predictor of a good wine. The box plots in fig.3 below illustrate the distribution of the variables according to good or poor wine quality. We can clearly see that we really do have a lot of variables to consider, and using graphs to select variables that have a noticeable effect on wine quality is far from easy. (B. Scibilia, 2011, [http://goo.gl/6RXK4K]).

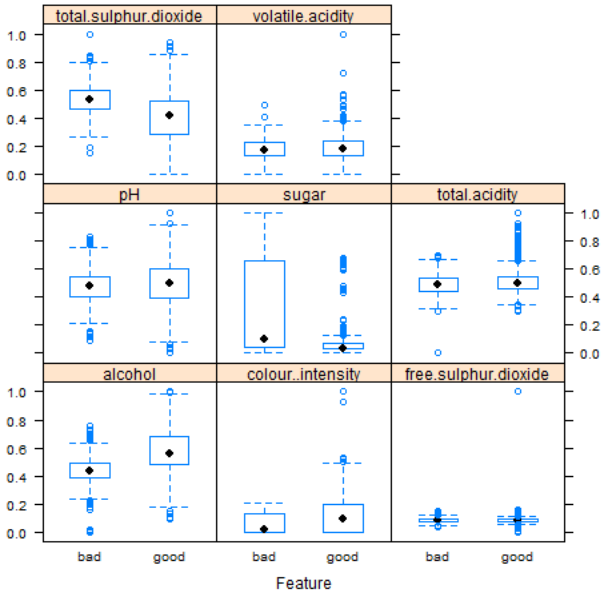

Fig. 3 Boxplots of parameters vs. Quality

2.3 Principal Component Analysis

In this stage before the beginning of regression analysis, we use a Principal Components (multivariate) Analysis to detect collinearity or correlation among the variables. Identifying variables that are highly collinear—which can make one of the variables almost redundant in some cases—can help us select the best possible binary logistic regression model and not only. In Fig.4 we can see these directions of the variables.

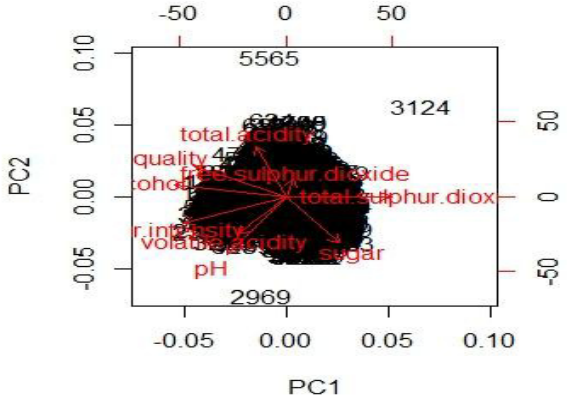

Fig.4 Loading Plot

It is more than obvious that quality and alcohol go very strongly together in the same direction, this mean "collinearity", on the contrary the sugar and total sulphur dioxide pointing on the other side mean that decreasing these parameter's values increase the first two.

Let see the result of the correlation matrix:

Principal Components Analysis

Call: principal(r = data2, nfactors = 1)

Standardized loadings (pattern matrix) based upon correlation matrix

	PC1	h2	u2	com
quality	0.73	0.5262	0.47	1
alcohol	0.88	0.7816	0.22	1
pH	0.47	0.2214	0.78	1
total.acidity	0.24	0.0589	0.94	1
volatile.acidity	0.38	0.1463	0.85	1
sugar	-0.42	0.1729	0.83	1
colour.intensity	0.81	0.6535	0.35	1
free.sulphur.dioxide	-0.06	0.0039	1.00	1
total.sulphur.dioxide	-0.82	0.6796	0.32	1

	PC1
SS loadings	3.24
Proportion Var	0.36

Mean item complexity = 1

Test of the hypothesis that 1 component is sufficient.

The root mean square of the residuals (RMSR) is 0.12
with the empirical chi square 6188.16 with prob < 0

Fit based upon off diagonal values = 0.84

These results confirm the above visual observations. Fig. 5 shows the correlations visually.

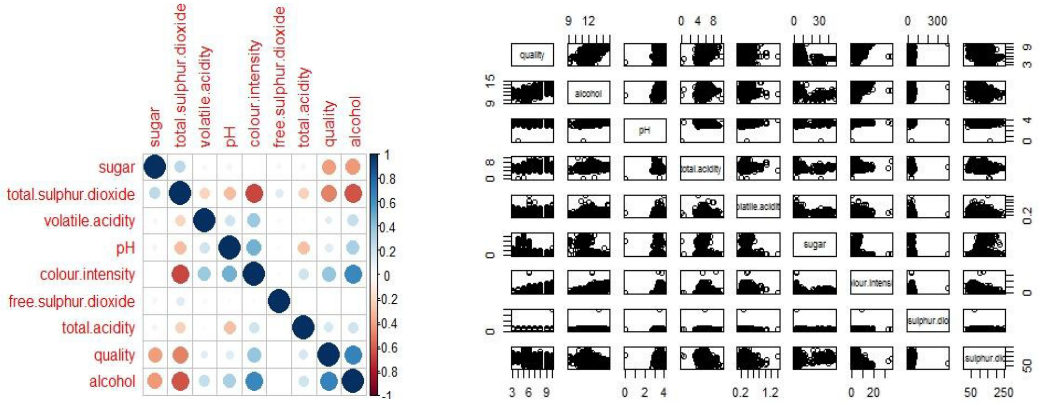

Fig.5 (a) correlation matrix (b) scatter-plot

3. Linear Regression Model

As mentioned earlier our first try to fit a model and predict quality score in our datasets is the Linear Regression. Showing that our target is quality we make linear regression with the variable "quality" vs. all others variables. If we are following the standard practice in regression analysis , we would realize that the R-squared is very low , so we consider separate models for white , red, and rose models .

The first result is from the white dataset and follows the other two.

3.1 Linear regression in white wines

If the response variable quality is assumed continuous the R code regression model is:
lm(formula = quality ~ alcohol + pH + total.acidity + volatile.acidity + sugar + colour.intensity + free.sulphur.dioxide + total.sulphur.dioxide)

and gives the following residuals and coefficients:

Residuals:

Min	1Q	Median	3Q	Max
-5.8042	-0.8522	0.0384	0.9747	3.9723

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-8.283097	0.904147	-9.161	< 2e-16 ***
alcohol	1.738027	0.061785	28.130	< 2e-16 ***
pH	-1.515943	0.157979	-9.596	< 2e-16 ***
total.acidity	0.017164	0.045188	0.380	0.704108
volatile.acidity	2.600841	0.310154	8.386	< 2e-16 ***
sugar	-0.036244	0.004074	-8.897	< 2e-16 ***
colour.intensity	-1.039690	0.314754	-3.303	0.000971 ***
free.sulphur.dioxide	-0.003613	0.005099	-0.709	0.478572
total.sulphur.dioxide	-0.006145	0.001422	-4.322	1.61e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.309 on 2306 degrees of freedom
Multiple R-squared: 0.4473, Adjusted R-squared: 0.4454
F-statistic: 233.3 on 8 and 2306 DF, p-value: < 2.2e-16

As we saw from the results the most critical coefficient the R-squared is 0.4473. The residuals comport very well converging to 0 with mean: -1.244592e-16.

3.2 Linear regression in red wines

If the response variable quality is assumed continuous the R code regression model is:
lm(formula = quality ~ alcohol + pH + total.acidity + volatile.acidity + sugar + colour.intensity + free.sulphur.dioxide + total.sulphur.dioxide)

and gives the following residuals and coefficients:
Residuals:

Min	1Q	Median	3Q	Max
-5.1216	-0.6101	0.0965	0.7649	3.6459

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-6.8361402	1.1740570	-5.823	6.45e-09 ***
alcohol	0.6417206	0.0373932	17.161	< 2e-16 ***
pH	1.7724375	0.2539107	6.981	3.66e-12 ***
total.acidity	0.1897624	0.0487405	3.893	0.000101 ***
volatile.acidity	-1.8084946	0.2464971	-7.337	2.86e-13 ***
sugar	-0.0163250	0.0024969	-6.538	7.39e-11 ***
colour.intensity	0.1305356	0.0101403	12.873	< 2e-16 ***
free.sulphur.dioxide	0.0067910	0.0024893	2.728	0.006411 **
total.sulphur.dioxide	-0.0156688	0.0009347	-16.764	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.153 on 2786 degrees of freedom
Multiple R-squared: 0.6402, Adjusted R-squared: 0.6391
F-statistic: 619.6 on 8 and 2786 DF, p-value: < 2.2e-16

The R-squared is **0.6402** and residuals comport very well with mean: -3.249814e-18 Nevertheless all the parameters have smaller than 0.01 values in P-significance.

3.3 Linear regression in roze wines

If the response variable quality is assumed continuous the R code regression model is:
lm(formula = quality ~ alcohol + pH + total.acidity + volatile.acidity + sugar + colour.intensity + free.sulphur.dioxide + total.sulphur.dioxide)

and gives the following residuals and coefficients:
Residuals:

Min	1Q	Median	3Q	Max
-3.0548	-1.2403	0.3824	0.9758	2.7585

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-4.615081	2.729274	-1.691	0.091488 .
alcohol	1.189249	0.135596	8.771	< 2e-16 ***
pH	-0.619562	0.597421	-1.037	0.300224
total.acidity	0.227841	0.102283	2.228	0.026369 *
volatile.acidity	-4.049047	0.967508	-4.185	3.39e-05 ***
sugar	-0.013273	0.010730	-1.237	0.216688
colour.intensity	-0.297197	0.134546	-2.209	0.027649 *
free.sulphur.dioxide	0.040575	0.011100	3.655	0.000285 ***
total.sulphur.dioxide	-0.015244	0.003737	-4.079	5.28e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.385 on 485 degrees of freedom
Multiple R-squared: 0.4062, Adjusted R-squared: 0.3964
F-statistic: 41.47 on 8 and 485 DF, p-value: < 2.2e-16

The behavior of linear regression in roze wine dataset is a little bit strange. Although has a good value in R-squared (**0.4062**) it is the minimum of all the three tests and the residuals have a not normal distribution although the mean is low (**8.128531e-17**).

Table 2 shows all the results gathered and fig.6 the residuals distributions.

Table 2. Gathered results from linear regression

	white	red	roze
Multiple R-squared	0.4473	0.6402	0.4062
Residuals mean	-1.244592e-16	-3.249814e-18	8.128531e-17

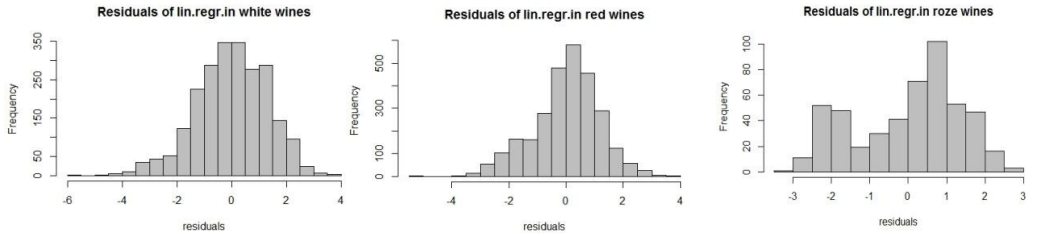

Fig. 6 Distributions of residuals / dataset

Obviously the second model in red wines has the better performance at all.

4. Logistic Regression Model

Knowing the results from linear regression model, we try to investigate another powerful model, the Logistic Regression in order to achieve better performance in prediction. Of course the main difference from the linear regression is that now we do not predict a score rather a category of quality: bad or good.

4.1 Logistic regression for white wines

If the response variable quality is assumed continuous the R code regression model is:

```
glm(formula = quality ~ alcohol + pH + total.acidity + volatile.acidity + sugar + colour.intensity + free.sulphur.dioxide + total.sulphur.dioxide, family = binomial(link = logit), data = dfwbin)
```

and gives the following residuals and coefficients:

```
Deviance Residuals:
   Min      1Q  Median      3Q     Max
-3.1552  -0.2781  0.4629  0.7470  2.6408
Coefficients:
                    Estimate   Std. Error  z value  Pr(>|z|)
(Intercept)        -9.141685   2.145349    -4.261   2.03e-05 ***
alcohol             1.882102   0.158114     11.903  < 2e-16 ***
pH                 -2.221449   0.363889    -6.105   1.03e-09 ***
total.acidity      -0.273557   0.099762    -2.742   0.00611 **
volatile.acidity    0.912013   0.620892     1.469   0.14187
sugar              -0.135114   0.015378    -8.786   < 2e-16 ***
```

colour.intensity	-1.076313	0.851296	-1.264	0.20611
free.sulphur.dioxide	0.026644	0.010176	2.618	0.00884 **
total.sulphur.dioxide	-0.026115	0.003031	-8.617	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2879.7 on 2314 degrees of freedom
Residual deviance: 2090.4 on 2306 degrees of freedom
AIC: 2108.4

Number of Fisher Scoring iterations: 6

Prediction table:

Pred.var.	bad	good
Bad	337	81
Good	389	1508

(337+1508)/2315

[1] **0.7969762**

So the final percentage of right prediction is about **80%**.

4.2 Logistic Regression Model in Red Wines

If the response variable quality is assumed continuous the R code regression model is:
glm(formula = quality ~ alcohol + pH + total.acidity + volatile.acidity + sugar + colour.intensity + free.sulphur.dioxide + total.sulphur.dioxide, family = binomial(link = logit), data = dfrbin)

and gives the following residuals and coefficients:

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.62285	0.03147	0.14301	0.42296	2.60591

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.084889	4.059249	-0.760	0.4473
alcohol	-0.078902	0.133750	-0.590	0.5552
pH	2.252629	0.876930	2.569	0.0102 *
total.acidity	0.223083	0.166008	1.344	0.1790
volatile.acidity	-5.863587	0.781134	-7.507	6.07e-14 ***
sugar	-0.122841	0.008730	-14.072	< 2e-16 ***
colour.intensity	0.494242	0.052370	9.438	< 2e-16 ***
free.sulphur.dioxide	0.005569	0.012983	0.429	0.6680
total.sulphur.dioxide	-0.028329	0.003004	-9.430	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2830.2 on 2794 degrees of freedom
Residual deviance: 1376.6 on 2786 degrees of freedom
AIC: 1394.6
Number of Fisher Scoring iterations: 7

Prediction table:

Var.pred.	bad	good
Bad	304	87
Good	267	2137

(304+2137)/2795
[1] **0.8733453**

The prediction percentage is about **87%**.

4.3 Logistic Regression in Roze Wines

If the response variable quality is assumed continuous the R code regression model is:
glm(formula = quality ~ alcohol + pH + total.acidity + volatile.acidity + sugar + colour.intensity + free.sulphur.dioxide + total.sulphur.dioxide, family = binomial(link = logit), data = dfzbin)

and gives the following residuals and coefficients:

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.0776	-0.8218	0.3298	0.7696	1.9827

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-25.809190	7.841920	-3.291	0.000998 ***
alcohol	2.215358	0.413234	5.361	8.27e-08 ***
pH	0.494436	1.262565	0.392	0.695345
total.acidity	0.250321	0.230497	1.086	0.277476
volatile.acidity	-3.158942	1.954637	-1.616	0.106067
sugar	-0.022269	0.020229	-1.101	0.270980
colour.intensity	-1.719609	0.662597	-2.595	0.009452 **
free.sulphur.dioxide	0.036977	0.019700	1.877	0.060520 .
total.sulphur.dioxide	-0.017534	0.007505	-2.336	0.019469 *

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
Null deviance: 620.73 on 493 degrees of freedom
Residual deviance: 453.33 on 485 degrees of freedom AIC: 471.33
Number of Fisher Scoring iterations: 6
Prediction table:

Var.Pred.	bad	good
bad	69	51
good	90	284

(69+284)/494
[1] **0.7145749**
So the percentage of right predictions is about **71.5 %**

5. Conclusions and future research proposals

The interest in wine quality score and especially predicting that score knowing physicochemical properties is increasing the recent years. With the arrival of the certification and the automating chemical recognition of properties, statistics and information technology are more than indispensables in the processing of wine. Of course the sensory test (or the human taste) is the final judge but nowadays knowing such big data of many years of productions and tasting results, it is possible to combine all these scientific sectors for the producer and the economy benefit.

This real case study with a large dataset of Greek wines was addressed by two main and classical DM methods or statistics regression models: the linear and the logistic. As we have seen the results of course favor the logistic model, because obviously the division is in two parts: good and bad, but the results with the linear especially in the red wine dataset show that a more accurate prediction even in a quality scale of 1 to 10 it is possible with good chances.

Table 3 gathers all the results

Table 3. Gathered results from linear and logistic regression

	white	red	roze
Linear	0.4473	0.6402	0.4062
Logistic	0.7969	0.8733	0.7145

The evolution in statistic modeling and the arising of new DM methods and models such Neural Networks, Support Vector Machines, Random Forest etc, give the opportunity in the near future to try testing our datasets with these methods, comparing the results and searching better performance.

On the other hand all these methods tested in this study could be used for any dataset of food or product viewing to achieve a better quality score. As we know these days with the economy every day more competitive and full of information data about any product, it is crucial to have as soon as possible high level data results for business decision making.

Acknowledgements

The authors dedicate this paper to the memory of Christina, Ioannis' wife, who showed so much courage in the whole of her life giving the main meaning in Ioannis' life. Also, thanks to the people and structure of the University of Macedonia for their support in this work .

References

Asuncion A. and Newman D., 2007, UCI Machine Learning Repository, University of California, Irvine, [http://www.ics.uci.edu/~mlearn/MLRepository.html].

Athanasiadis, I and Ioannides ,D. , 2014, Proceedings of the 27th Greek statistical conferences, A statistical analysis of big web market data structure using R and the direct e-commerce social phenomenon of producer to consumer transaction pp. 333-347

Banks, D. and Said, Y. , 2006, Data mining in electronic commerce, Statistical Science, 21(2), 234-246

Borle, S., Boatwright, P. and Kadane J. , 2006, The timing of bid placement and extent of multiple bidding: an empirical investigation using ebay online auctions, Statistical Science, 21(2), 194-205

Brynjolfsson, E., Hu, Y. and Smith, M. , 2003, Consumer surplus in the digital economy: Estimating the value of increased product variety at online booksellers, Management Sci., 49,1580-1596.

Chevalier, J. and Goolsbee, A. , 2003, Measuring prices and price competition online: Amazon.com and Barnes and Noble.com. Croall, J. (1997).

Cortez Paulo, António Cerdeirab, Fernando Almeidab, Telmo Matosb, José Reis (2009), Modeling wine preferences by data mining from physicochemical properties, Journal Decision Support Systems, Volume 47, Issue 4, November 2009, Pages 547–553

Dass, M. and Reddy, S. , 2006, Modeling on-line art auction dynamics using functional data analysis, Statistical Science, 21(2), 179-193

Dass, M. and Reddy, S. , 2008, An analysis of price dynamics, bidder networks and market structure in online art auctions, in book: Statistical methods in e-commerce research, Ed. W.Jank and G.Shmueli, John Wiley & Sons, Inc., 105-129

Ferrer J., MacCawley A., Maturana S., Toloza S., and Vera J.., 2008, An optimization approach for scheduling wine grape harvest operations. International Journal of Production Economics, 112(2):985–999.

Fienberg, S. , 2006, Privacy and confidentiality in an e-commerce world: data mining, data warehousing, matching and disclosure limitation, Statistical Science, 21(2), 143-154

Gesell, S.,1958, The Natural Economic Order, Revised edition. London: Peter Owen, (http://wikilivres.info/wiki/The_Natural_Economic_Order)

Ghose, A. and Sundararajan, A. ,2006, Evaluating pricing strategy using e-commerce data: evidence and estimation challenges, Statistical Science, 21(2), 131-142

Greene, W. H. ,2000, Econometric Analysis, 4th ed., Prentice-Hall, Upper Saddle River, NJ.

Guyon I. and Elisseeff A.,2003, An introduction to variable and feature selection. Journal of Machine Learning Research, 3(7–8):1157–1182.

Miller W.T. ,2013, Modeling Techniques in Predictive Analytics: Business Problems and Solutions with R, FT Press

P. K. Janert K.P. ,2011, Data Analysis with Open Source Tools, O'Reilly

Quandt, R. E. ,1964, Statistical discrimination among alternative hypotheses and some economic regularities, J. Regional Sci. 5 1–23.

Ramsay, J.O. and Silverman, B.W. ,2005, Functional Data Analysis, New York: Springer- Verlag

Siegel E. ,2013, Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die, Wiley

Simonoff, J.S. ,1996, Smoothing Methods in Statistics, New York: Springer-Verlag. Quantitative Marketing and Economics 1 203-222

Smith D. and Margolskee R.,2006, Making sense of taste. Scientific American, Special issue, 16(3):84–92.

Turban E., Sharda R., Aronson J., and King D.,2007, Business Intelligence, A Managerial Approach. Prentice-Hall

Witten H.I. and Frank E.,2005, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco, CA,2nd edition.