Group H - Assignment 4

This assignment was focused on predicting certain properties of a dataset with several wine characteristics. The dataset contained 13 different characteristics connected to the chemical composition of the wine and one column showing customer segment. Furthermore, linear regression, logistic regression, and principal component analysis are used to analyze the dataset. Afterward, the results of those different techniques are evaluated and compared.

1. Predicting the percentage of alcohol in wine with linear regression

Linear regression is a very effective tool to predict a certain characteristic on the basis of a set of other variables. In our case, the column containing a percentage volume of alcohol was used as a dependent variable and the other twelve chemical characteristics and customer segment were used as independent ones. Furthermore, Statsmodels package was used to execute the regressions and the standard output is added as Exhibit 1. The regression resulted in an adjusted R-squared of 58.7%, which indicates that the model explains 58.7% of the variability of the dependent variable. Since it exceeds the threshold of 20-30%, it is considered a well-fitted model. Moreover, the p-value for the F-statistic is far lower than 5%, which indicates that the regressors in the model are relevant to explain the dependent variable.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                Alcohol   R-squared:                       0.663
Model:                            OLS   Adj. R-squared:                  0.634
Method:                 Least Squares   F-statistic:                     22.89
Date:                Wed, 12 Dec 2018   Prob (F-statistic):           1.52e-31
Time:                        15:01:47   Log-Likelihood:                -118.20
No. Observations:                 178   AIC:                             266.4
Df Residuals:                     163   BIC:                             314.1
Df Model:                          14
Covariance Type:            nonrobust

                         coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------
Malic_Acid             0.0745      0.044      1.704      0.090      -0.012       0.161
Ash                   -0.2858      0.213     -1.342      0.181      -0.706       0.135
Ash_Alcanity           0.0044      0.018      0.242      0.809      -0.031       0.040
Magnesium             -0.0001      0.003     -0.033      0.974      -0.006       0.006
Total_Phenols          0.0988      0.125      0.793      0.429      -0.147       0.345
Flavanoids            -0.0505      0.115     -0.438      0.662      -0.278       0.177
Nonflavanoid_Phenols  -0.0373      0.405     -0.092      0.927      -0.837       0.762
Proanthocyanins       -0.0671      0.091     -0.736      0.463      -0.247       0.113
Color_Intensity        0.1243      0.030      4.142      0.000       0.065       0.184
Hue                    0.3522      0.263      1.339      0.183      -0.167       0.872
OD280                  0.0308      0.109      0.282      0.778      -0.185       0.246
Proline                0.0002      0.000      0.726      0.469      -0.000       0.001
Customer_Segment_2    -1.1190      0.198     -5.657      0.000      -1.510      -0.728
Customer_Segment_3    -0.7246      0.298     -2.434      0.016      -1.313      -0.137
const                 12.8898      0.652     19.767      0.000      11.602      14.177
```

Exhibit 1 - Output of the linear regression with Alcohol as dependent and chemical and customer properties as independent variables

2. Logistic regression

Logistic regression is used in cases, where the predicted value is binary. In this case, the strength of the wine was predicted after setting an arbitrary line of using top 25% amounts of alcohol as “strong” wines. Logistic regression allows us to assume non-constant effects among independent variables of the regression as well as non-standard error terms. The results of the regression with all the covariates used can be seen in Exhibit 3 and 4.

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Malic_Acid	0.7963	0.3299	2.4139	0.0158	0.1497	1.4429
Ash	-0.4046	1.4830	-0.2728	0.7850	-3.3111	2.5020
Ash_Alcanity	-0.1211	0.1261	-0.9605	0.3368	-0.3683	0.1260
Magnesium	-0.0174	0.0246	-0.7090	0.4783	-0.0655	0.0307
Total_Phenols	1.7637	1.0456	1.6868	0.0916	-0.2856	3.8130
Flavanoids	0.1078	1.0286	0.1048	0.9165	-1.9081	2.1237
Nonflavanoid_Phenols	0.5048	3.2619	0.1548	0.8770	-5.8884	6.8980
Proanthocyanins	-1.2858	0.8099	-1.5876	0.1124	-2.8731	0.3016
Color_Intensity	0.7868	0.2686	2.9289	0.0034	0.2603	1.3133
Hue	8.5526	2.6106	3.2761	0.0011	3.4359	13.6692
OD280	1.1993	0.8031	1.4933	0.1354	-0.3748	2.7734
Proline	0.0006	0.0016	0.3521	0.7247	-0.0026	0.0037
intercept	-16.7840	5.5746	-3.0108	0.0026	-27.7099	-5.8580
Customer_Segment_2	-2.4306	1.5222	-1.5968	0.1103	-5.4141	0.5528
Customer_Segment_3	1.6249	2.4489	0.6635	0.5070	-3.1749	6.4246

Exhibit 3 - Coefficients of the logistic regression with binary variable Strong as dependent

Model:	Logit	Pseudo R-squared:	0.459
Dependent Variable:	Strength	AIC:	143.3607
Date:	2018-12-12 14:02	BIC:	191.0875
No. Observations:	178	Log-Likelihood:	-56.680
Df Model:	14	LL-Null:	-104.74
Df Residuals:	163	LLR p-value:	2.6115e-14
Converged:	1.0000	Scale:	1.0000
No. Iterations:	8.0000		

Exhibit 4 - Statistics for Exhibit 3

3. Logistic regression with PCA component

The Principal Component Analysis was used to reduce the number of variables in our dataset to 4 components that together explained more than 75% of the variance of the dataset. Logistic regression was run with the new components but the results were less accurate when compared with the non-PCA version described in the previous point.

Model:	Logit	Pseudo R-squared:	0.079
Dependent Variable:	Strength	AIC:	200.8556
Date:	2018-12-12 14:06	BIC:	213.5827
No. Observations:	178	Log-Likelihood:	-96.428
Df Model:	3	LL-Null:	-104.74
Df Residuals:	174	LLR p-value:	0.00084224
Converged:	1.0000	Scale:	1.0000
No. Iterations:	6.0000		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
0	-0.2583	0.0746	-3.4631	0.0005	-0.4045	-0.1121
1	-0.6680	0.1163	-5.7448	0.0000	-0.8958	-0.4401
2	-0.2012	0.1447	-1.3905	0.1644	-0.4849	0.0824
3	-0.0598	0.1803	-0.3319	0.7400	-0.4132	0.2935

Exhibit 6 - The outcomes of logistic regression with PCA components

4. Comparison logistic regression with vs. without PCA component

As the comparison of the two previously run models, gain and lift curves and confusion matrix are shown in this part of the text. As can be seen from Exhibit 7 regression with PCA gives an accurate prediction for the first 10%-20% of the sample but then the performance of the model drops significantly. Following, Gain curve in Exhibit 8 shows a steeper increase in prediction in the first 30% with which it is possible to predict almost 80% of cases of the strong wines. At last, Exhibit 9 gives clear evidence of the supreme accuracy of the model without the usage of PCA.

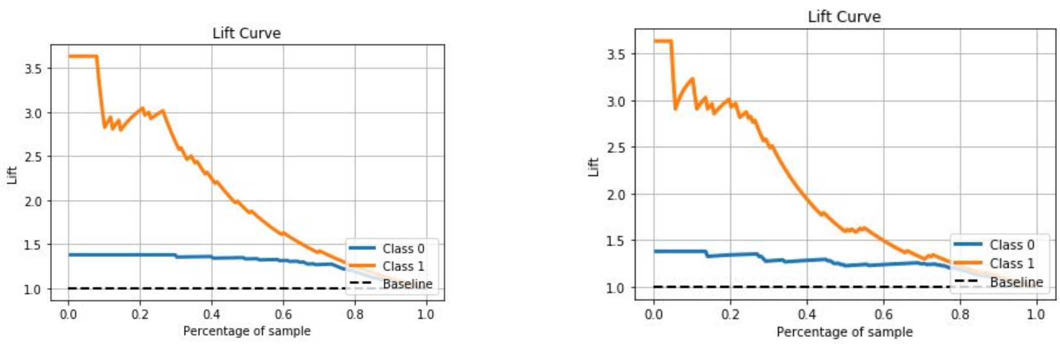

Exhibit 7 - Lift Curve for a logistic model without PCA (left) and with PCA (right)

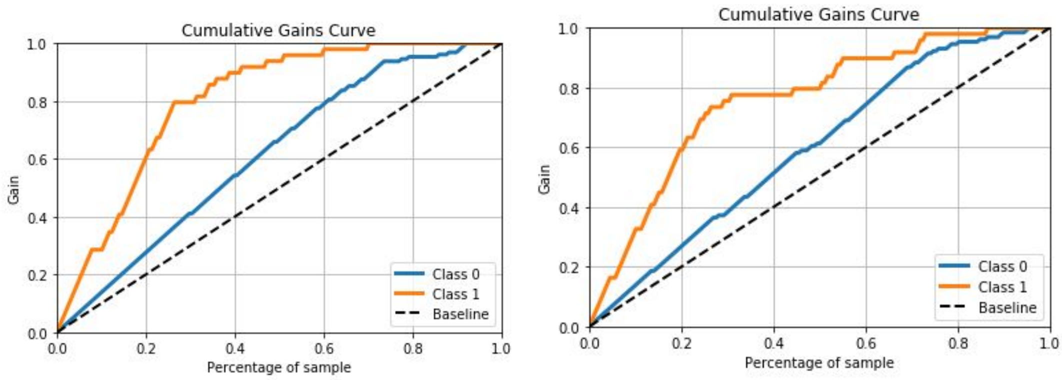

Exhibit 8 - Gain Curve for a logistic model without PCA (left) and with PCA (right)

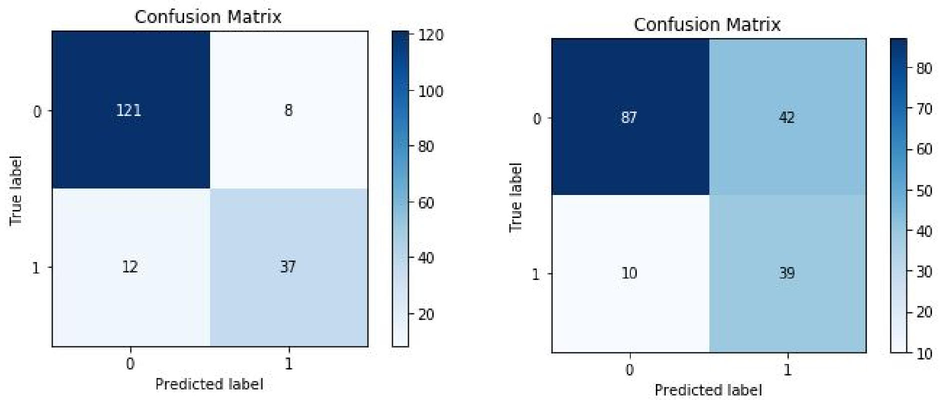

Exhibit 9 - Confusion Matrix for a logistic model without PCA (left) and with PCA (right)