Модель предсказания качества вин

Отчет

Зачем мы это делали?

Нашей **целью** было научиться предсказывать качество вина.

Задача возникла в связи с тем, что возможность предсказать качество вина напрямую влияет на издержки компании.

Сейчас чтобы произвести вино высокого качества нам необходимо перебрать больше 100 вариаций основных параметров (таких как процент алкоголя, кислотность итд). Разработка, производство и распространение одного сорта низкого качества обходится компании примерно в 50 000$.

Таким образом модель позволит сократить более половины заведомо успешных комбинаций тем самым сократив затраты компании на тестирование новых сортов вин на 2 500 000$. Помимо этого компания сможет доставлять качественные продукты на рынок быстрее.

На каких данных мы это делали?

Для модели в течении полугода мы собирали информацию о разных сортах выпускаемых нашей компанией и их потребительских оценках. Большинство параметров было предоставлено нашей лабораторией. Пример наших данных приведен ниже.

	type	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	white	7.0	0.27	0.36	20.7	0.045	45.0	170.0	1.0010	3.00	0.45	8.8	6
1	white	6.3	0.30	0.34	1.6	0.049	14.0	132.0	0.9940	3.30	0.49	9.5	6
2	white	8.1	0.28	0.40	6.9	0.050	30.0	97.0	0.9951	3.26	0.44	10.1	6
3	white	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6
4	white	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6

Какие модели мы использовали?

Для предсказания оценки мы использовали только логистическую регрессию со стандартными параметрами. Это было сделано исключительно из-за нехватки времени и компетенций.

```
model = LogisticRegression()
```

```
model.fit( X_train , y_train )
```

```
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                   intercept_scaling=1, l1_ratio=None, max_iter=100,
                   multi_class='warn', n_jobs=None, penalty='l2',
                   random_state=None, solver='warn', tol=0.0001, verbose=0,
                   warm_start=False)
```

Что у нас получилось?

На выходе мы получили модель которая дает точность порядка 50%.

К плюсам модели можно отнести то, что она не переобучена и ее построение заняло очень мало времени.

К минусам - достаточно низкое качество и точность.

```
# Score the model
print (model.score( X_train , y_train ) , model.score( X_test , y_test ))
```

```
0.5323017408123791 0.5127610208816705
```

Что мы будем делать дальше?

Так как модель еще окончательно не готово и ее качество оставляет желать лучшего, мы планируем:

1. Поработать с признаками и добавить новые
2. Сбалансировать выборку для лучшего результата
3. Разделить задачу на несколько подзадач (определение недопустимо низкого качества и предсказание оценки для сортов приемлемого качества)
4. Попробовать другие классификаторы

Спасибо за внимание