УДК 004.85

Использование библиотеки Scikit-learn и Orange для создания модели предсказания стоимости вина и оценки качество модели

Черкашин Александр Михайлович
Приамурский государственный университет имени Шолом-Алейхема
Студент

Аннотация
Целью является использования программы по предсказание будущие стоимость на вина. В работе использовано библиотека Scikit-learn и составления схемы работы (Orange) с данными для создания и обучение модели и предсказание стоимость вина и оценка модели. В результате обученная модель будет предсказывать стоимость вина.
Ключевые слова: Python, Scikit-learn, Orange-canvas, Линейная регрессия.

Using the Scikit-learn library and Orange to create a model for predicting the cost of wine and assessing the quality of the model

Cherkashin Alexander Mihailovich
Sholom-Aleichem Priamursky State University
Student

Abstract
The goal is to use a program to predict the future value of wines. The work used the Scikit-learn library and mapping work (Orange) with data to create and train a model and predict the cost of wine and evaluate the model. As a result, the trained model will predict the cost of wine.
Keywords: Python, Scikit-learn, Orange-canvas, Linear regression.

Научный руководитель:
Баженов Руслан Иванович,
Приамурский государственный университет имени Шолом-Алейхема,
к.п.н., доцент, зав. кафедрой информационных систем, математики и правовой информатики

1 Введение

1.1 Актуальность исследования

Данная статья описывает возможность написание скрипта и составления схемы работы с данными для обучения моделей вина и предсказание стоимость вина и оценка модели.

1.2 Цель исследования

Целью исследования является написание скрипта на языке Python и составление схемы работы с данными по обучение моделей и предсказание стоимость вина и оценка модели.

1.3 Обзор исследований

Р. И. Новиков, Е. Г. Романова рассмотрели алгоритм логистической регрессии разработана программа на языке программирования Python [1]. А. Е. Вертинская провел сравнение методов машинного обучения по таким критериям, как качество предсказаний, загрузка процессора, использование памяти и диска, а также разработка методики выбора подходящего алгоритма согласно требованиям задачи [2]. С. В. Федотов представил краткий обзор методов ансамблирования моделей машинного обучения стекинга, бэггинга, бустинга [3]. М. М. Постников, Б. В. Добров показал возможность применение нейронные сети с использованием переноса обучения для специальной размеченной коллекции [4]. П. И. Морозов, Е. Г. Романова исследовал применение решающих деревьев [5]. Мазуренко В. А. показал возможность предсказание колебаний цен на биткойн [6].

2. Результаты и обсуждение

2.1 Исходные данные

Исходные данные взяты из источника автора курса https://stepik.org/course/73952 (дата обращения 2021-11-16) в каталоге https://stepik.org/lesson/411692/step/1 (дата обращения 2021-11-16) в ссылке на презентации, слайд 93 (Анализ цены вина – практическое решение) в слайде указано ссылка на скачивания.

Атрибуты:

Meta — Мета, не используется для обучение нейронный сети.

Target — Цель, значение который что предсказать.

Feature — Исходные данные, указывает что обучать.

Skip — Пропускать, не используется в определении столбец таблицы.

Столбец таблицы: Имя столбец — Описания (Атрибут)

- Идентификатор

Year — Год хранения вина (Meta)

Price — Стоимость вина (Target)

WinterRain — Уровень осадков зимой (Feature)

AGST — Средне сезонная температура (Feature)

HarvestRain — Уровень осадков летом (Feature)

Age — Возраст вина (Feature)

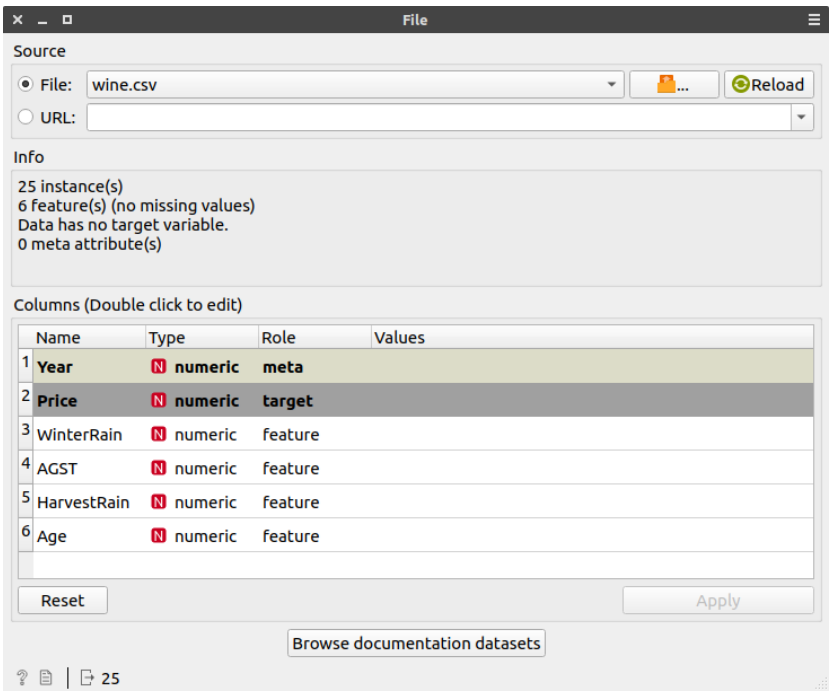

Рисунок 2.1. Загрузка файла (File) wine.csv, и заданным Role для столбец данными (Orange-canvas)

Для модели искусственный интеллект значение WinterRain, AGST, HarvestRain, Age определен данные для обучение нейронный сеть, Price определен проверяющий ответ, который должно получиться (рис 2.1). Данная модель используется для обучения с учителем (линейная регрессия). Используется приложение Orange-canvas-core [7] http://orange.biolab.si/ (https://orangedatamining.com/) для составления схемы работы с данными (рис 2.2).

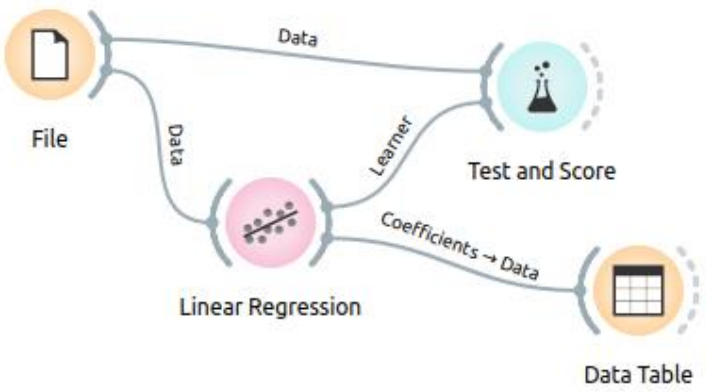

Рисунок 2.2. Схема построение с использование Orange-canvas

Исходный код 2.1 написан на языке Python, использует библиотеки matplotlib [8] для визуализация данных (построение графики) [9]. Библиотека scikit-learn (sklearn) для машинного обучения [10]. Pandas библиотека для обработки и анализа данных [11].

Листинг 2.1 Исходный код

```
#!/usr/bin/python3
# -*- coding: utf-8 -*-

import matplotlib.pyplot as plt
from sklearn import linear_model
from sklearn.metrics import mean_squared_error, r2_score, mean_absolute_error

import pandas as pd
df = pd.read_csv("wine.csv")
df_train = df[["WinterRain", "AGST", "HarvestRain", "Age"]]
df_target = pd.DataFrame({"Price": df["Price"]})
# Выбираем модель линейная регрессия
regr = linear_model.LinearRegression()
# Обучаем модель
regr.fit(df_train, df_target)
df_test = regr.predict(df_train)
print("Coefficients: \n", regr.coef_)
m_mse = mean_squared_error(df_target, df_test)
print("MSE: %.3f" % m_mse)
print("RMSE: %.3f" % m_mse ** (1/2))
print("MAE: %.3f" % mean_absolute_error(df_target, df_test))
print("R2: %.3f" % r2_score(df_target, df_test))

plt.scatter(df["Year"], df_target, color="black")
plt.plot(df["Year"], df_test, color="blue", linewidth=3)

plt.show()
```

В результате выполнение программы вывод

```
Coefficients:
 [[ 0.00107551  0.60720935 -0.00397153  0.02393083]]
MSE: 0.070
RMSE: 0.264
MAE: 0.226
R2: 0.829
```

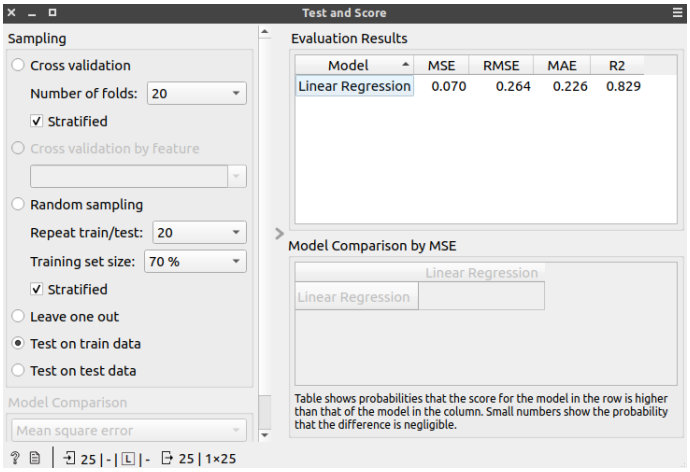

Рисунок 2.3. Test and Score результат (Orange-canvas)

В результате получаем значение R2: 0.829.

Строка 4 - 8 (листинг 2.1) Импортируем библиотеку, matplotlib, scikit-learn, pandas.

Строка 9 (листинг 2.1) импортируем данные wine.csv.

Строка 10 (листинг 2.1) Выбираем столбец как исходный данные (Feature).

Строка 11 (листинг 2.1) Выбираем столбец для целевой данные (правильный ответ) (Target).

Строка 13 (листинг 2.1) Выбираем модель линейная регрессия.

Строка 15 (листинг 2.1) Обучаем модель.

Строка 16 (листинг 2.1) Вводим модель, для получение результата насколько модель подстроилась к исходным данным для получения ответа. Используется для сравнения ответом (данные черный точки) и модель полученный ответ (синий линий) (рис 2.4).

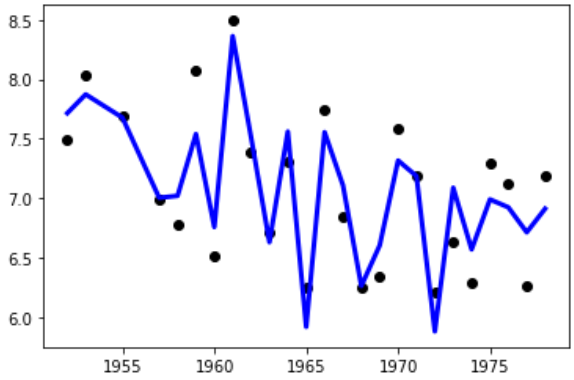

Рисунок 2.4. График вывода, обученная модель

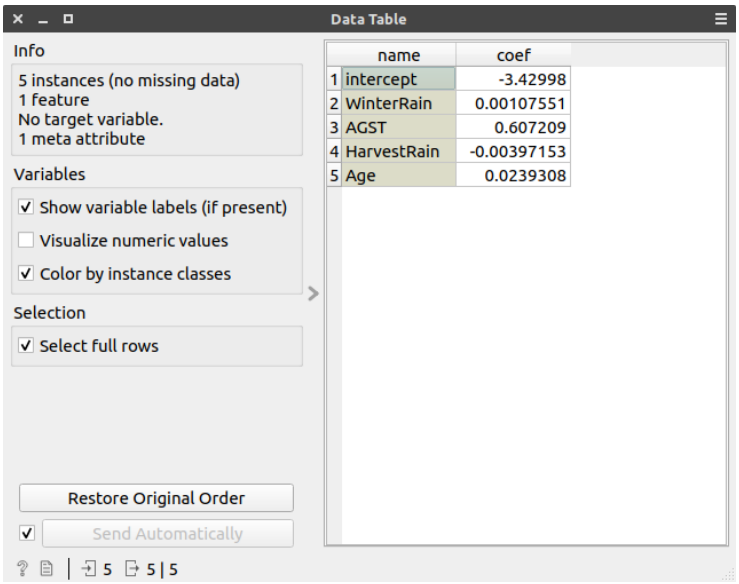

Рисунок 2.5. Результаты полученный коэффициенты coef

Строка 17 (листинг 2.1) выводит значение коэффициенты coef_ (рис 2.5), содержит массив, определенный Feature для предсказания каждый отдельный цель [12, 13].

В данном коэффициент в 2 ключ массив значения 0.60720935 это столбец AGST указывает признак сильного влияния на цель.

Строка 18 и 19 (листинг 2.1) mean_squared_error (MSE) — функция метрика вычисляет среднюю квадратичная отклонения для оценки риска ожидаемому потери [14].

Низкий значения указывает признак низкий уровень риска в модели (качественный модель низкий уровень ошибки) используемый оценки MSE.

Строка 20 (листинг 2.1) RMES это извлеченный квадратный корень на MSE [15].

Строка 21 (листинг 2.1) mean_absolute_error (MAE) — Вычисляет среднюю абсолютную ошибку (метрика риска).

Строка 22 (листинг 2.1) r2_score вычисляет коэффициент детерминант определения долю дисперсии, для определения независимыми переменными в модели.

Высокие значения указывает наличие качественный модель (R2).

Строка 24 и 25 (листинг 2.1) Функция построение графики. Горизонтально значение это Year (Год), вертикальное значение — Черные точки это Price (Цена), синие линии (Тестируем модель) это ввод значение Feature в обученный модель полученный перемена Test (predict(df_train)) (Строка 16 листинг 2.2) (Рис 2.4).

Строка 27 (листинг 2.1) - Отобразить график.

3 Выводы

В данной статье был описан скрипт по обучению модели для предсказания стоимости вина, по метрике R2 показывается наличие качественный модели. В исследовании применялись два способа работы. 1 Способ через Orange-canvas. 2. Способ через написание скрипта Python. В результате получаем одинаковый результат. Были использованы различные метрики для оценки MSE, RMSE, MAE, R2, и коэффициенты (Coef) для оценки конкретного исходного столбца, а также выполнено сравнение целевых данных и обученной модели в наглядно построенном графике.

Библиографический список

1. Новиков Р. И., Романова Е. Г. Классификация объектов по признакам с помощью алгоритма логистической регрессии //Математическое и компьютерное моделирование естественно-научных и социальных проблем: материалы XIII Меж. 2019. С. 230.
2. Вертинская А. Е. Методика сравнения методов машинного обучения в зависимости от различных параметров задачи: магистерская диссертация: дис. БГУ, ФПМИ, Кафедра дискретной математики и алгоритмики, 2020.
3. Федотов С. В. Сопоставление характеристик ансамблевых методов машинного обучения в задачах оценки оттока клиентов на примере сети фитнес-клубов //Современные технологии: актуальные вопросы, достижения и инновации. 2020. С. 23-26.

4. Постников М. М., Добров Б. В. Представление новостных сюжетов с помощью событийных фотографий //Аналитика и управление данными в областях с интенсивным использованием данных. 2017. С. 438-445.
5. Морозов П. И., Романова Е. Г. Применение метода решающих деревьев //Математическое и компьютерное моделирование естественно-научных и социальных проблем: материалы XIII Меж. – 1994. – Т. 62. – С. 71.
6. Мазуренко В. А. и др. Прогнозирование дневных изменений цен на Биткойн с помощью методов интеллектуального анализа текста. – 2018.
7. orange-canvas-core · PyPI // PyPI URL: https://pypi.org/project/orange-canvas-core/ (дата обращения: 2022-01-15).
8. matplotlib · PyPI // PyPI URL: https://pypi.org/project/matplotlib/ (дата обращения: 2022-01-15).
9. Matplotlib — Visualization with Python // matplotlib URL: https://matplotlib.org (дата обращения: 2022-01-15).
10. scikit-learn: machine learning in Python — scikit-learn 1.0.2 documentation // scikit-learn URL: https://scikit-learn.org/ (дата обращения: 2022-01-15).
11. pandas - Python Data Analysis Library // pandas URL: https://pandas.pydata.org/ (дата обращения: 2022-01-15).
12. sklearn.linear_model.LinearRegression — scikit-learn 1.0.2 documentation // scikit-learn URL: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html (дата обращения: 2022-01-15).
13. scikit learn - What is target in Python's sklearn coef_ output? - Stack Overflow // stackoverflow URL: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html (дата обращения: 2022-01-15).
14. 3.3. Metrics and scoring: quantifying the quality of predictions — scikit-learn 1.0.2 documentation // scikit-learn URL: https://scikit-learn.org/stable/modules/model_evaluation.html#mean-squared-error (дата обращения: 2022-01-15).
15. python - rmse cross validation using sklearn - Stack Overflow // stackoverflow URL: https://stackoverflow.com/questions/69432869/rmse-cross-validation-using-sklearn (дата обращения: 2022-01-15).

Приложения

Таблица 1. Исходные данные о вине (файл: wine.csv).

#	Year	Price	WinterRain	AGST	HarvestRain	Age
0	1952	7.4950	600	17.1167	160	31
1	1953	8.0393	690	16.7333	80	30
2	1955	7.6858	502	17.1500	130	28
3	1957	6.9845	420	16.1333	110	26

4	1958	6.7772	582	16.4167	187	25
5	1959	8.0757	485	17.4833	187	24
6	1960	6.5188	763	16.4167	290	23
7	1961	8.4937	830	17.3333	38	22
8	1962	7.3880	697	16.3000	52	21
9	1963	6.7127	608	15.7167	155	20
10	1964	7.3094	402	17.2667	96	19
11	1965	6.2518	602	15.3667	267	18
12	1966	7.7443	819	16.5333	86	17
13	1967	6.8398	714	16.2333	118	16
14	1968	6.2435	610	16.2000	292	15
15	1969	6.3459	575	16.5500	244	14
16	1970	7.5883	622	16.6667	89	13
17	1971	7.1934	551	16.7667	112	12
18	1972	6.2049	536	14.9833	158	11
19	1973	6.6367	376	17.0667	123	10
20	1974	6.2941	574	16.3000	184	9
21	1975	7.2920	572	16.9500	171	8
22	1976	7.1211	418	17.6500	247	7
23	1977	6.2587	821	15.5833	87	6
24	1978	7.1860	763	15.8167	51	5

Рисунок 2.6. Исходный код (файл wine_data.png)

Листинг 3.1. SHA 256:

```
77b28e26997fdd4f083dd0a6629a9a8d4b18ca778436d63e0c32c51a76e8342a
```

Команда convert входит в состав программы ImageMagick (https://imagemagick.org) для обработки изображения. Используем для преобразования изображения (рис 2.6) в архив.

В Листинг 3.2. использовался ОС Ubuntu.

В Листинг 3.3. использовался ОС Windows 10.

Файл (рис 2.6) содержится 1 бит глубина цвета и 3 компонента (RGB) размер изображения 232×236.

Процедура преобразования:

1. В файл изображения (рис 2.6) преобразовать в архив
2. Удалить лишний 4 байта.
3. Проверить целостность (листинг 3.1).
4. Распаковать файл архив.

Листинг 3.2. Под Unix

```
1 convert -depth 1 wine_data.png rgb:wine_data.cpio.xz
2 truncate -s -4 wine_data.cpio.xz
3 shasum -a 256 wine_data.cpio.xz
4 xz -cd wine_data.cpio.xz | cpio -id
```

Листинг 3.1, строка 1. используется программа ImageMagick для чтения изображения файл (рис 2.6) формата png и преобразования в файл.

Листинг 3.1, строка 2. используется утилита для изменения размера файла, входящий пакет GNU core utilities (https://www.gnu.org/software/coreutils/) вычитается 4 байта.

Листинг 3.1, строка 3. используется утилита для проверки целостность файла (SHA 256) хеш должно совпадать в листинг 3.1.

Листинг 3.1, строка 4. используется утилита (пакет xz-utils https://tukaani.org/xz/) и cpio (https://www.gnu.org/software/cpio/manual/cpio.html) для распаковки файла wine_data.cpio.xz.

Листинг 3.3. Под Windows

```
1 convert -depth 1 wine_data.png rgb:wine_data.cpio.xz
2 wmic datafile where Name='<Полный путь>\\wine_data.cpio.xz' get Size
3 FSUTIL file seteof wine_data.cpio.xz <Размер файла>
4 certUtil -hashfile wine_data.cpio.xz SHA256
```

В Листинг 3.3, строка 1. используется программа ImageMagick для чтения изображения формата png и преобразования в файл (рис 2.6).

В Листинг 3.3, строка 2. Используется утилита для получения размер файла. Указывается полный путь, путь должен быть \\ а не \.

В Листинг 3.3, строка 3. Используется утилита для изменения размер файла в данном случае указывается текущий размер файла и вычитать 4 байта (<размер файла> - 4).

В Листинг 3.3, строка 4. Используется утилита, используется для проверки целостность файла (SHA 256) хеш должно совпадать в листинг 3.1.

Файл wine_data.cpio.xz распаковывается любым архиватором используя например 7-zip (https://www.7-zip.org/) или (https://tukaani.org/xz/).