Avito BI contest

Василий Рубцов

25 февраля 2017

Задача 2

Метрика f1score

	1 predicted	0 predicted
1 actual	TP	FN
0 actual	FP	TN

Таблица Confusion matrix

Задача 2

Метрика f1score

	1 predicted	0 predicted
1 actual	TP	FN
0 actual	FP	TN

Таблица Confusion matrix

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Задача 2

Метрика f1score

	1 predicted	0 predicted
1 actual	TP	FN
0 actual	FP	TN

Таблица Confusion matrix

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{F}_1 = \frac{2}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}} = \frac{2\text{TP}}{2\text{TP} + \text{FN} + \text{FP}}$$

Задача 2

Метрика f1score

	1 predicted	0 predicted
1 actual	TP	FN
0 actual	FP	TN

Таблица Confusion matrix

Пусть TP, TN, FP, FN будут обозначать соответсвующие доли от всей выборки.

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{F}_1 = \frac{2}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}} = \frac{2\text{TP}}{2\text{TP} + \text{FN} + \text{FP}}$$

Задача 2

Метрика f1score

	1 predicted	0 predicted
1 actual	TP	FN
0 actual	FP	TN

Таблица Confusion matrix

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{F}_1 = \frac{2}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}} = \frac{2\text{TP}}{2\text{TP} + \text{FN} + \text{FP}}$$

Пусть TP, TN, FP, FN будут обозначать соответсвующие доли от всей выборки.
Если пометить все объекты единицей:

$$\text{F}_1 = \frac{2\text{TP}}{2\text{TP} + \text{FN} + 0}, \quad \text{TP} + \text{FP} = 1$$

Задача 2

Метрика f1score

	1 predicted	0 predicted
1 actual	TP	FN
0 actual	FP	TN

Таблица Confusion matrix

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{F}_1 = \frac{2}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}} = \frac{2\text{TP}}{2\text{TP} + \text{FN} + \text{FP}}$$

Пусть TP, TN, FP, FN будут обозначать соответсвующие доли от всей выборки.
Если пометить все объекты единицей:

$$\text{F}_1 = \frac{2\text{TP}}{2\text{TP} + \text{FN} + 0}, \quad \text{TP} + \text{FP} = 1$$

Узнали долю единиц. Обозначим ее за x.

Задача 2

Метрика f1score

	1 predicted	0 predicted
1 actual	TP	FN
0 actual	FP	TN

Таблица Confusion matrix

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{F}_1 = \frac{2}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}} = \frac{2\text{TP}}{2\text{TP} + \text{FN} + \text{FP}}$$

Пусть TP, TN, FP, FN будут обозначать соответсвующие доли от всей выборки.
Если пометить все объекты единицей:

$$\text{F}_1 = \frac{2\text{TP}}{2\text{TP} + \text{FN} + 0}, \quad \text{TP} + \text{FP} = 1$$

Узнали долю единиц. Обозначим ее за x.
За y обозначим долю единиц в сабмите.
Тогда:

$$\begin{cases} \text{TP} + \text{TN} + \text{FP} + \text{FN} = 1 \\ \text{TP} + \text{FN} = \text{x} \\ \text{TP} + \text{FP} = \text{y} \\ \frac{2\text{TP}}{2\text{TP}+\text{FN}+\text{FP}} = \text{F}_1 \end{cases}$$

Задача 3

Данные и метрика

Предсказание количества просмотров объявлений

$$\text{RMSLE} = \sqrt{\frac{\sum_{i=1}^{n}(\log(y_i + 1) - \log(\hat{y}_i + 1))^2}{n}}$$

Задача 3

Данные и метрика

Предсказание количества просмотров объявлений

$$\text{RMSLE} = \sqrt{\frac{\sum_{i=1}^{n}(\log(y_i + 1) - \log(\hat{y}_i + 1))^2}{n}}$$

	start_time	title	price	owner_type	category	subcategory	param1	param2	param3	region	item_views
0	2016-12-27 10:38:04	Сандали фирмы Crocs	800	Private	Личные вещи	Детская одежда и обувь	Для мальчиков	Обувь	> 36	Москва	27
1	2016-12-27 15:23:55	Бутсы футбольные Reebok	2000	Private	Личные вещи	Детская одежда и обувь	Для мальчиков	Обувь	> 36	Омская область	9
2	2016-12-28 19:34:15	Nike hypervenom Бутсы	600	Private	Личные вещи	Детская одежда и обувь	Для мальчиков	Обувь	> 36	Санкт-Петербург	105
3	2016-12-26 10:26:02	Сапоги	150	Private	Личные вещи	Детская одежда и обувь	Для мальчиков	Обувь	> 36	Тульская область	28

Рис. Обучающая выборка

Задача 3

Обработка признаков

1) start_time → количество секунд от начала дня
2) title → нормализация слов (pymorphy2) → one-hot 5000 самых популярных слов (sklearn.feature_extraction.text.CountVectorizer)
3) price
4) Label encoding категориальных признаков

Задача 3

Обработка признаков

1) start_time → количество секунд от начала дня
2) title → нормализация слов (pymorphy2) → one-hot 5000 самых популярных слов (sklearn.feature_extraction.text.CountVectorizer)
3) price
4) Label encoding категориальных признаков
5) Частотность категориальных признаков

Задача 3

Обработка признаков

1) start_time → количество секунд от начала дня
2) title → нормализация слов (pymorphy2) → one-hot 5000 самых популярных слов (sklearn.feature_extraction.text.CountVectorizer)
3) price
4) Label encoding категориальных признаков
5) Частотность категориальных признаков
6) Среднее значение ключевого признака внутри каждой категории для param1

Задача 3

Фильтрация шумовых объектов

```
for train, valid in cv:
    model.fit(X[train], y[train])
    p = model.predict(X[valid])
    error[valid] = abs(p - y[valid])
X = X[error < threshold]
y = y[error < threshold]
```

Задача 3

Фильтрация шумовых объектов

```
for train, valid in cv:
    model.fit(X[train], y[train])
    p = model.predict(X[valid])
    error[valid] = abs(p - y[valid])
X = X[error < threshold]
y = y[error < threshold]
```

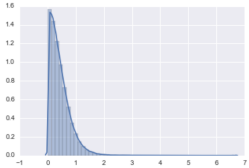

Рис. Распределение ошибок

Задача 3

Модель

- Несколько xgboost с разной глубиной (10, 12, 20)
  0.54561
- Блендинг
  0.54385