

SASCOMPETITIONS:

Прогнозирование вероятности невозврата
кредита

Соболева Дарья
МГУ им. Ломоносова

15 декабря 2017

Задача

Разработка модели кредитного скоринга, основанной на анализе данных о кредитном поведении клиентов, полученных из внешних источников.

Данные

ID	SK_DATE_ DECISION	NUM_ SOURCE	f1	f2	...	fn	DEF
24368	20150901	1					0
24368	20150901	3					0
24368	20150901	3					0
24489	20150901	1					1
...

f1 ... fn -- атрибуты кредита из кредитной истории

Даты

- SK_DATE_DECISION
- DTIME_CREDIT
- DTIME_CREDIT_ENDDATE
- DTIME_CREDIT_ENDDATE_FACT

(Greater_date - Lower_date).days

- DTIME_CREDIT_UPDATE

(Curr_date - DTIME_CREDIT_UPDATE).days

Текст

- CREDIT_CURRENCY

value_counts()

- TEXT_PAYMENT_DISCIPLINE

value_counts(), len()

Статистики. Вещественные признаки.

- CREDIT_ACTIVE
- CREDIT_SUM_TYPE

```
data.groupby('ID')[f].value_counts()
```

- AMT_ANNUITY
- DATES
- TEXT
- ...

```
data.groupby('ID')[f].apply(func)
```

func ∈ min, max, mean, median

Статистики. Категориальные признаки.

- CREDIT_FACILITY
- CREDIT_TYPE
- CREDIT_ACTIVE
- CREDIT_SUM_TYPE

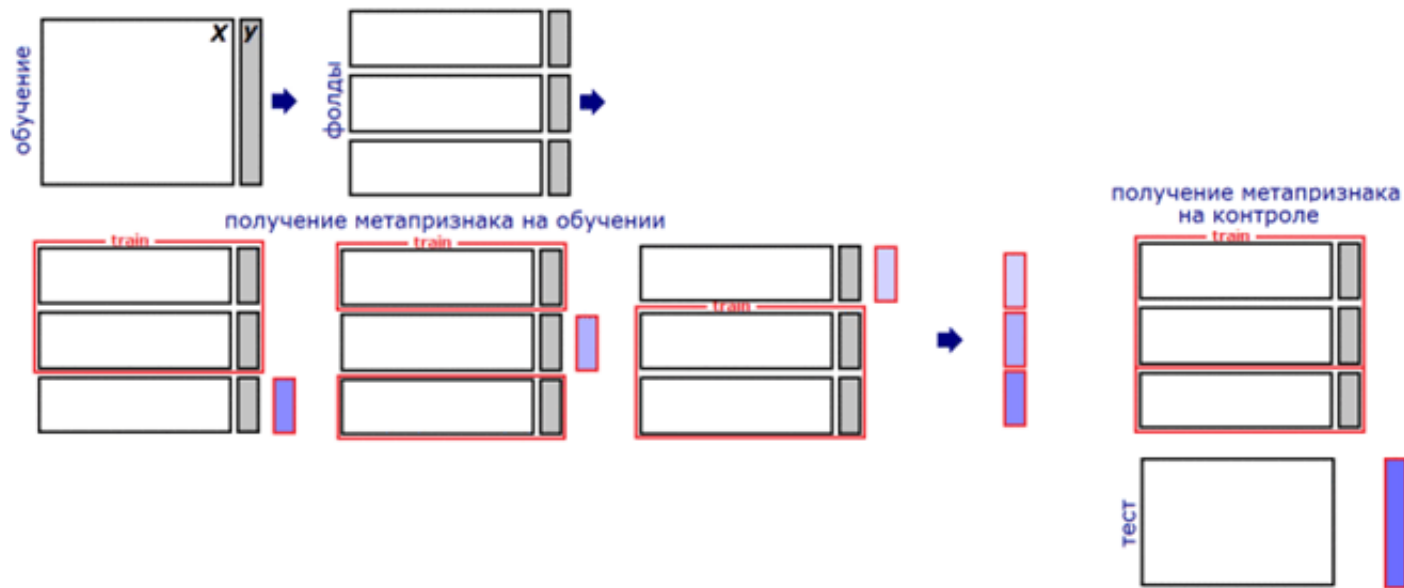
```
data[f].groupby('ID').apply(func)
```

```
func ∈ min_freq_val, max_freq_val
```

Статистики по таргету

```
data.groupby(f)['DEF'].mean()
```


Модель



Gradient Boosting (SAS), LightGBM (Python)
Logistic Regression (SAS, Python).

Спасибо за внимание

Вопросы