Прогноз принятия законопроектов

От Константина Тихонова

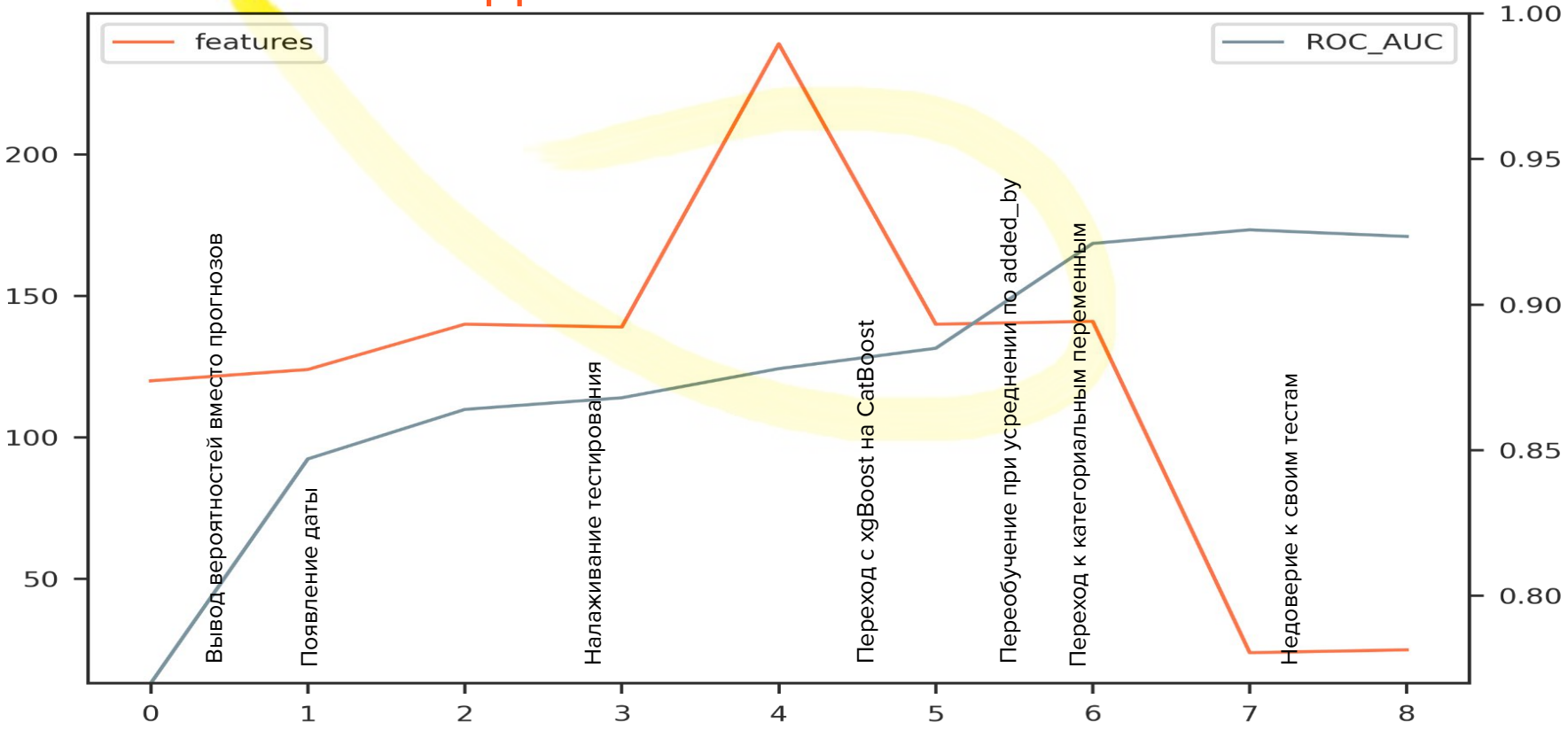
Что за модель?

1. 25 ± 1-2 переменные
2. CatBoost
3. Минимальное использование текстов законопроектов и ria_reports_main.csv
4. Ключевые столбцы в regulations.csv:
   a. added_by
   b. responsible
   c. publication_date
   d. developer

Список всех переменных

added_by
developer
year
responsible
responsible_count
views_num
regulatory_impact
mineco_solution
year_end
month
npa_type
act_objectives
text_len
co_dev_mean
dislikes_num
проект
act_significance
problem_addressed
text
relations_regulated_by_act
persons_affected_by_act
okved_list
is_regionally_signigicant
act_changes_controlling_activities
co_developer_bool

Как менялась модель



Тестирование **vs Kaggle leaderboard**

1. Все данные, по которым проверялся ответ, имеют текст. НПА с текстом имеют несколько другое распределение, чем без текста. Поэтому моя тестовая выборка состояла только из данных с текстом.

2. Необходима более или менее надежная система тестирования, с помощью которой можно сравнивать модели: отбирать переменные, настраивать параметры и т.д.

3. Тестовая выборка на Kaggle - 2000, данных с текстом и ответом, на которых можно тестироваться - 7384. Поэтому внутренний тест должен быть намного более точным, чем результат в Kaggle и именно на него надо ориентироваться.

4. Сабмит по всей выборке.

Пример тестов

Seed 6100 With same_person responsible_count added_by_count mean: 0.9004592451188526 With same_person without проект mean: 0.9002500442961365

Seed 6100 With same_person responsible_count mean: 0.9004684687492048 With same_person added_by_count mean: 0.8994416600489498

Seed 6100 With same_person mean: 0.9009906512855516 Without mean: 0.9004776493667135

Seed 6100 With responsible_count mean: 0.9010515423107959 With added_by_count mean: 0.9004826031569817

With ['year', 'month'] as categorical features mean: 0.902169056237123 With ['month', 'year_end'] as categorical mean: 0.9021170452319177