Направление 02.03.02 —
Фундаментальная информатика и
информационные технологии

Использование нейронных сетей для прогнозирования вероятности заинтересованности пользователя контентом в рамках веб сайта

Выпускная квалификационная работа на степень
бакалавра

Игнатова Анастасия Алексеевна, 4 курс 8 группа

Научный руководитель: М. И. Чердынцева, к. т. н.

Постановка задачи

1. Выполнить программную реализацию моделей для предсказания поведения пользователя веб-сайта;
2. Экспериментально проверить качество решений применением моделей к тестовому набору данных;
3. Провести сравнительный анализ и интерпретацию результатов исследований

Задача от яндекса

	sample_id	item	publisher	user	topic_0	topic_1	topic_2	topic_3	topic_4	weight_0	weight_1	weight_2	weight_3	weight_4
0	1009109	1716	349	1053	362	397	430	287	431	54	54	51	26	13
1	1009110	1707	202	254	150	73	356	212	482	29	7	5	5	4
2	1009111	1592	520	1524	397	287	356	330	281	95	46	6	5	3
3	1009112	1541	82	2994	397	287	102	323	356	93	77	25	7	4
4	1009113	52	520	936	201	283	618	249	617	35	33	30	11	9

тестовый набор данных

	sample_id	item	publisher	user	topic_0	topic_1	topic_2	topic_3	topic_4	weight_0	weight_1	weight_2	weight_3	weight_4	target
0	0	531	147	2925	411	477	618	249	460	27	18	9	8	7	0
1	1	1574	260	2981	212	287	382	302	51	27	11	2	1	0	0
2	2	940	394	1230	145	150	212	170	174	7	6	6	5	5	0
3	3	52	520	2597	201	283	618	249	617	35	33	30	11	9	1
4	4	766	55	1680	362	150	477	305	388	51	15	13	10	9	1

тренировочный набор данных

Градиентный бустинг

$$Loss = \sum_{i=1}^{N} \Psi(y_i, F(x_i)) \quad \textbf{1}$$

$$(\beta_m, \alpha_m) = \arg min_{\beta,\alpha} \sum_{i=1}^{N} \Psi(y_i, F_{m-1}(x_i) + \beta h(x_i, \alpha)) \quad \textbf{2}$$

$$F_m(x) = F_{m-1}(x) + \beta_m h(x, \alpha_m) \quad \textbf{3}$$

Градиентный бустинг

$$\alpha_m = \arg min_{\beta,\alpha} \sum_{i=1}^{N} [-g_m(x_i) - \beta \mathrm{h}(x_i, \alpha)]^2 \quad \textbf{4}$$

$$g_m(x_i) = [\frac{\partial \Psi(y_i, \mathrm{F}(x_i))}{\partial\, \mathrm{F}(x_i)}]_{\mathrm{F}(x)=\mathrm{F}_{m-1}(x)}, i = 1..N \quad \textbf{5}$$

$$\rho_m = \arg min_{\rho} \sum_{i=1}^{N} \Psi(y_i, \mathrm{F}_{m-1}(x_i) + \rho \mathrm{h}(x_i, \alpha_m)) \quad \textbf{6}$$

Расширение набора данных

$$CTR = \frac{total\ number\ of\ clicks}{total\ number\ of\ impressions} * 100\%$$

	sample_id	item	publisher	user	topic_0	topic_1	topic_2	topic_3	topic_4	weight_0	weight_1	weight_2	weight_3	weight_4	target	ctr
0	0	531	147	2925	411	477	618	249	460	27	18	9	8	7	0	153005
1	1	1574	260	2981	212	287	382	302	51	27	11	2	1	0	0	137838
2	2	940	394	1230	145	150	212	170	174	7	6	6	5	5	0	218391
3	3	52	520	2597	201	283	618	249	617	35	33	30	11	9	1	190272
4	4	766	55	1680	362	150	477	305	388	51	15	13	10	9	1	312749

Конструирование моделей

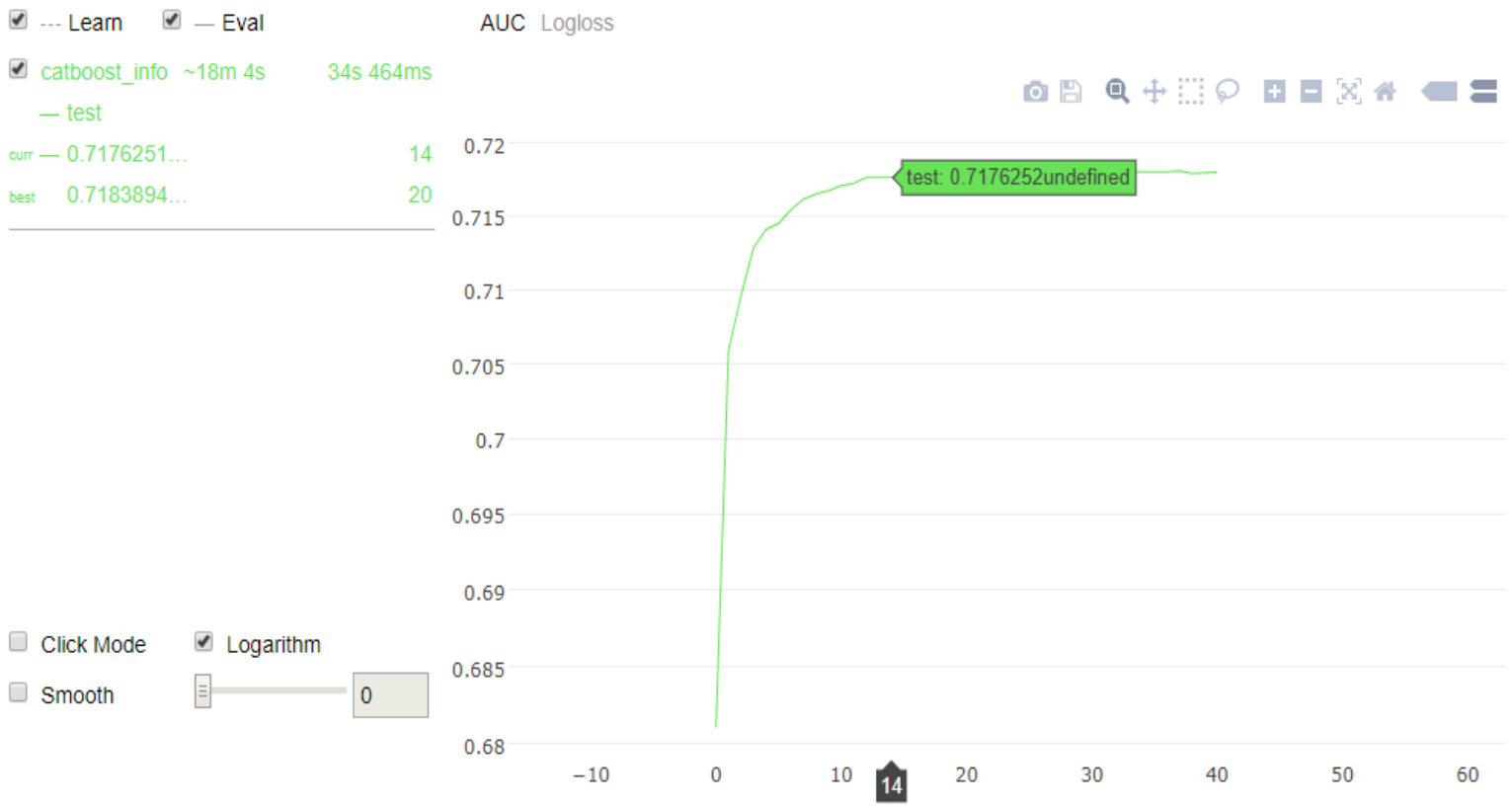

График обучения модели

Анализ результата

	Модель №1	**Модель №2**	**Модель №1 с CTR**	**Модель №2 с CTR**
Время работы	1 мин 52 сек	16 мин 33 сек	4 мин 55 сек	24 мин 5 сек
MSLp1E(Mean squared logarithmic error)	0.3301	0.3295	0.3248	0.3254
MAE(Mean Absolute Error)	0.7987	0.7976	0.7884	0.79
MSE(Mean Squared Error)	0.6534	0.6520	0.6436	0.6442

Полученные результаты

1. В рамках выпускной квалификационной работы решена задача, для которой были построены рекомендательные системы с использованием метода градиентного бустинга.

2. Сформированы и обучены две модели. Для обучения использовались изначальные и расширенные данные Расширение происходило за счет введения параметра CTR.

3. В результате проведённого проекта, было выявлено, что реализованная модель способна решать данную задачу с вероятностью 81,7%.