ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Выявление мошенников на торговой площадке Авито

Бунак Алексей Валерьевич

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Постановка задачи

1. провести разведочный анализ данных
2. выполнить препроцессинг (предобаботку)
3. сравнить модели по точности классификации и скорости обучения
4. разработать нейронную сеть
5. сравнить нейронную сеть с моделями машинного обучения

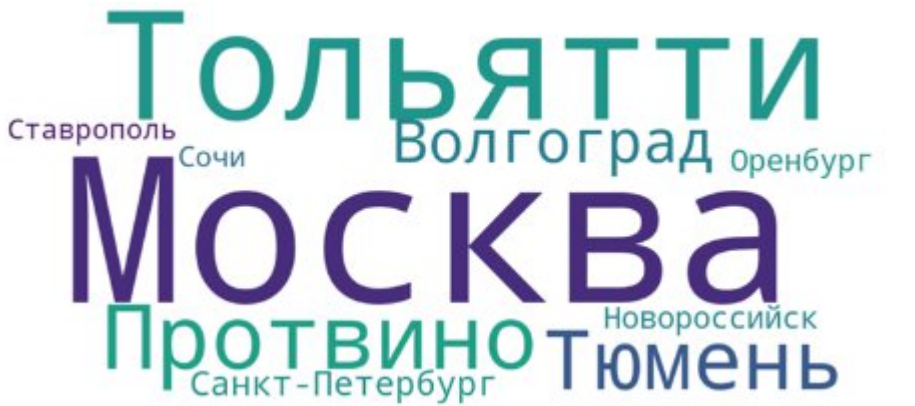
Дополнительные задачи

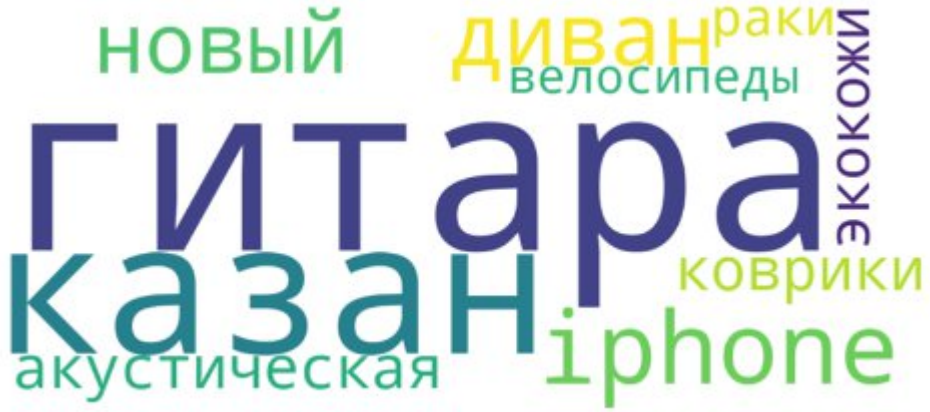
Разработать веб-приложение

Выложить на GitHub

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Анализ данных

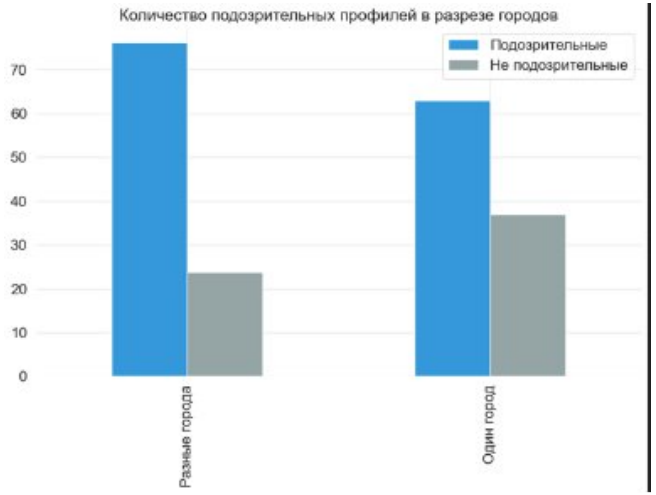


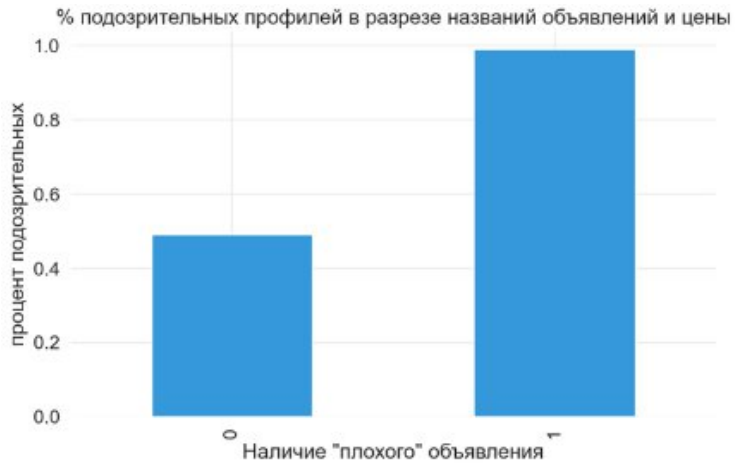


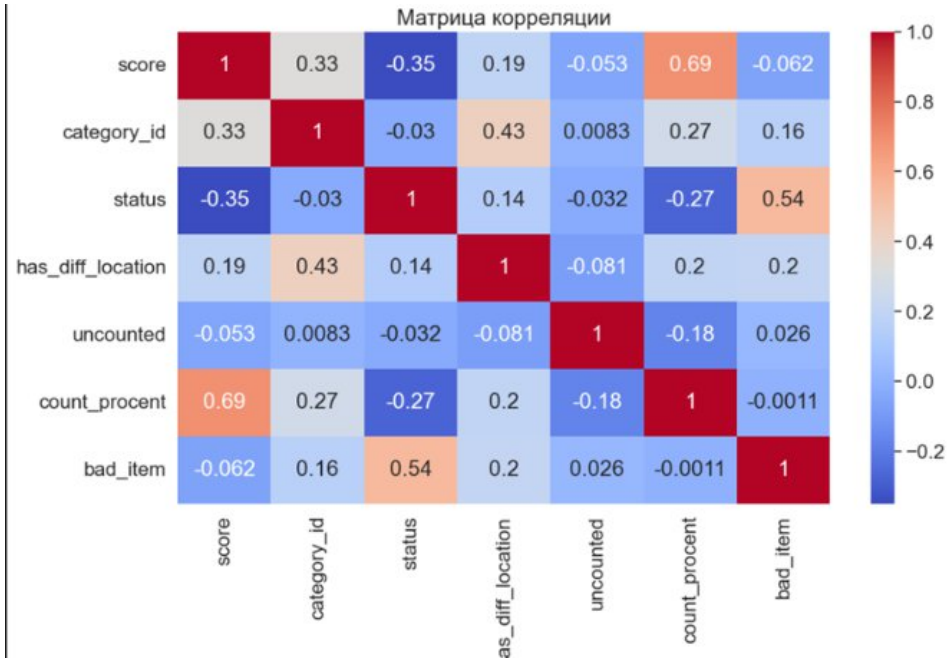
avito_blacklist	Обзор	Структура	Поиск	Вставить	Очистить	Удалить	1 753	MyISAM	utf8mb4_general_ci	240.5 КиБ		
avito_bots	Обзор	Структура	Поиск	Вставить	Очистить	Удалить	3	InnoDB	utf8mb4_general_ci	48.0 КиБ		
avito_bots_access	Обзор	Структура	Поиск	Вставить	Очистить	Удалить	4	MyISAM	utf8mb4_general_ci	4.1 КиБ		
avito_categories	Обзор	Структура	Поиск	Вставить	Очистить	Удалить	50	MyISAM	utf8mb4_general_ci	4.8 КиБ		
avito_chats	Обзор	Структура	Поиск	Вставить	Очистить	Удалить	614 818	MyISAM	utf8mb4_general_ci	311.0 МБ		
avito_chats_users	Обзор	Структура	Поиск	Вставить	Очистить	Удалить	2 012 696	MyISAM	utf8mb4_general_ci	354.4 МБ		
avito_cities	Обзор	Структура	Поиск	Вставить	Очистить	Удалить	2 602	MyISAM	utf8mb4_general_ci	209.2 КиБ		
avito_emails	Обзор	Структура	Поиск	Вставить	Очистить	Удалить	3 310	MyISAM	utf8mb4_general_ci	841.0 КиБ		
avito_items	Обзор	Структура	Поиск	Вставить	Очистить	Удалить	1 304 019	MyISAM	utf8mb4_general_ci	419.6 МБ		
avito_messages	Обзор	Структура	Поиск	Вставить	Очистить	Удалить	~4 774 927	InnoDB	utf8mb4_general_ci	3.2 ГиБ		
avito_messages_content	Обзор	Структура	Поиск	Вставить	Очистить	Удалить	275 810	MyISAM	utf8mb4_general_ci	145.4 МБ		
avito_messages_errors	Обзор	Структура	Поиск	Вставить	Очистить	Удалить	648	MyISAM	utf8mb4_general_ci	71.6 КиБ		
avito_phones	Обзор	Структура	Поиск	Вставить	Очистить	Удалить	88 737	MyISAM	utf8mb4_general_ci	17.1 МБ		
avito_profiles	Обзор	Структура	Поиск	Вставить	Очистить	Удалить	537 263	MyISAM	utf8mb4_general_ci	118.9 МБ		
avito_profiles_changes	Обзор	Структура	Поиск	Вставить	Очистить	Удалить	0	MyISAM	utf8mb4_general_ci	1.0 КиБ		
avito_rating	Обзор	Структура	Поиск	Вставить	Очистить	Удалить	1 989	MyISAM	utf8mb4_general_ci	159.0 КиБ		
avito_reviews	Обзор	Структура	Поиск	Вставить	Очистить	Удалить	18 739	MyISAM	utf8mb4_general_ci	9.9 МБ		
avito_socials	Обзор	Структура	Поиск	Вставить	Очистить	Удалить	1 645	MyISAM	utf8mb4_general_ci	376.2 КиБ		
avito_stats	Обзор	Структура	Поиск	Вставить	Очистить	Удалить	455 249	MyISAM	utf8mb4_general_ci	47.3 МБ		
avito_users	Обзор	Структура	Поиск	Вставить	Очистить	Удалить	688 359	MyISAM	utf8mb4_general_ci	127.6 МБ		
crm_accounts	Обзор	Структура	Поиск	Вставить	Очистить	Удалить	1 261	MyISAM	utf8mb4_general_ci	149.5 КиБ		
crm_balance_history	Обзор	Структура	Поиск	Вставить	Очистить	Удалить	0	MyISAM	utf8mb4_general_ci	1.0 КиБ		
crm_chats	Обзор	Структура	Поиск	Вставить	Очистить	Удалить	80	MyISAM	utf8mb4_general_ci	8.3 КиБ		
crm_paid_functions	Обзор	Структура	Поиск	Вставить	Очистить	Удалить	12	MyISAM	utf8mb4_general_ci	6.4 КиБ		
crm_settings	Обзор	Структура	Поиск	Вставить	Очистить	Удалить	1 363	MyISAM	utf8mb4_general_ci	153.1 КиБ		
crm_users	Обзор	Структура	Поиск	Вставить	Очистить	Удалить	5 670	MyISAM	utf8mb4_general_ci	2.4 МБ		
tg_messages	Обзор	Структура	Поиск	Вставить	Очистить	Удалить	~116 314	InnoDB	utf8mb4_general_ci	33.3 МБ		
28 таблиц	**Всего**						**~10 907 321**	**InnoDB**	**utf8mb4_general_ci**	**4.7 ГиБ**		

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Препроцессинг







Появились зависимости

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Сравнение алгоритмов

Метод	Точность	После кросс-валидации	Скорость
Линейная регрессия	89.43%	85.68%	1.4сек
Многослойный персептрон (MLP)	87.8%	85.06%	24.5сек
Метод опорных векторов	87.8%	85.06%	1м 30сек
Метод k-ближайших соседей	84.55%	86.08%	1.6сек
Деревья решений	88.62%	85.86%	1.8сек
Случайный лес	**91.06%**	**87.72%**	1м 3сек
Градиентный бустинг	87.8%	85.07%	3.1сек
Наивный байесовский алгоритм	89.43%	83.24%	1.2сек
Метод стохастического градиента	76.42%	77.92%	**1.1сек**
Нейронная сеть	**89.43%**		27.2 сек

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Заключение

Задача РЕШЕНА

Причины относительно низкого результата прогноза (можно улучшить):

- Очень малая выборка (53 тысячи строк из 1.3 миллионов)
- Человеческий фактор (bad_item настраивает человек)
- Малое количество признаков (я не имею доступа ко всем данным сайта Авито)
- Упрощение модели (для демонстрации работы мне пришлось убрать часть информации, которая будет оцениваться в реальном проекте)

ЦЕНТР
ДОПОЛНИТЕЛЬНОГО
ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

do.bmstu.ru