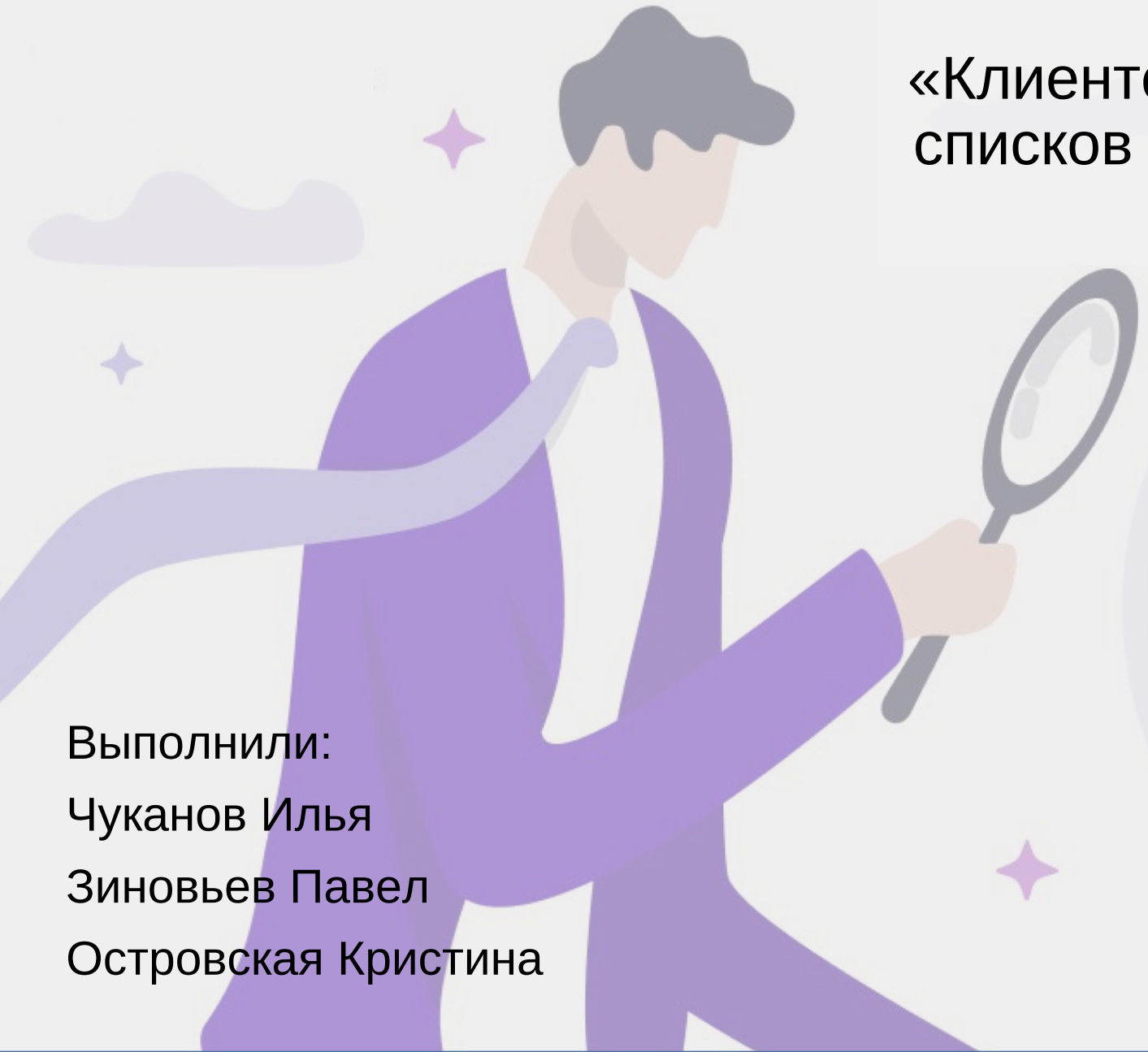
Групповой проект на тему

«Клиентская аналитика - Оптимизация списков клиентов на коммуникацию на основе истории контактов»

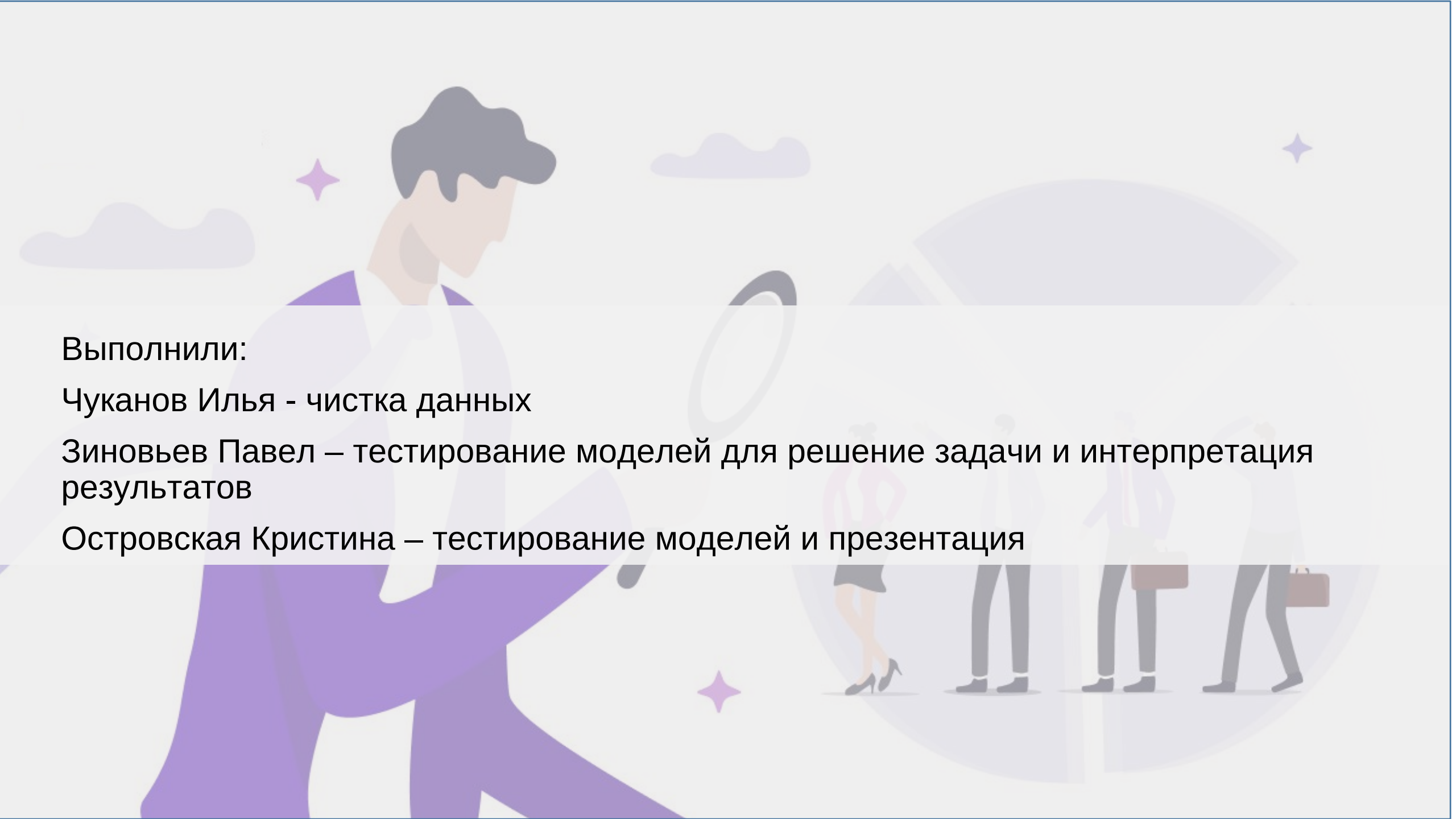


Выполнили:

Чуканов Илья

Зиновьев Павел

Островская Кристина



Выполнили:

Чуканов Илья - чистка данных

Зиновьев Павел – тестирование моделей для решение задачи и интерпретация результатов

Островская Кристина – тестирование моделей и презентация

Цели работы

Выделение целевой аудитории

Избежание затрат на взаимодействие с пользователем, который вероятно не заинтересован в предложении

Оптимизация маркетинга в компании

Задачи

Чистка данных

Применить несколько моделей для предсказания

Выделить выборку из заинтересованных клиентов для последующего взаимодействия

данные на входе

Датасет с информацией о клиентах и их откликах для e-mail и sms рассылок

данные на выходе

Аналитические модели, предсказывающие заинтересованность клиентов в продукте (количество откликов)

Этапы реализации проекта – чистка данных

- Удаление строк с пропущенными значениями: age, lifetime, gender
- До чистки данных - 985477 строк, после 733165

	Column Name	Zero Count	Unique Count	Zero Percentage	Null Count	Unique Percentage	Null Percentage
0	ID	0	985477	0.00	0	100.00	0.00
1	Age	0	63	0.00	66958	0.01	6.79
2	Ind_Household	640250	2	64.97	0	0.00	0.00
3	Age_group	578936	4	58.75	0	0.00	0.00
4	District	53206	56	5.40	0	0.01	0.00
5	Region	382905	6	38.85	0	0.00	0.00
6	Lifetime	2110	40	0.21	12608	0.00	1.28
7	Income	0	51	0.00	0	0.01	0.00
8	Segment	379739	4	38.53	0	0.00	0.00
9	Ind_deposit	797999	2	80.98	0	0.00	0.00
10	Ind_email	965956	2	98.02	0	0.00	0.00
11	Ind_phone	961490	2	97.57	0	0.00	0.00
12	Ind_salary	916207	2	92.97	0	0.00	0.00
13	trans_6_month	0	42505	0.00	0	4.31	0.00
14	trans_9_month	0	50868	0.00	0	5.16	0.00
15	trans_12_month	0	57702	0.00	0	5.86	0.00
16	amont_trans	0	44	0.00	0	0.00	0.00
17	amont_day_from	0	31	0.00	0	0.00	0.00
18	trans_3_month	0	65926	0.00	0	6.69	0.00
19	Gender	538741	3	54.67	0	0.00	0.00

Этапы реализации проекта – кодировка данных

trans_6_month	trans_9_month	trans_12_month	amont_trans	amont_day_from	trans_3_month
0.509187	0.402909	0.448284	0.041667	0.34375	0.417452
0.593804	0.540197	0.586768	0.208333	0.18750	0.621387
0.509628	0.446867	0.412410	0.041667	0.15625	0.575128
0.674348	0.668199	0.713072	0.020833	0.50000	0.626156
0.345506	0.335473	0.451276	0.083333	0.62500	0.267821

Стандартизация числовых признаков:

Кодировка категориальных признаков:

Ind_deposit_No	Ind_deposit_Yes	Ind_email_No	Ind_email_Yes	Ind_phone_No	Ind_phone_Yes	Ind_salary_No	Ind_salary_Yes
1	0	0	1	0	1	1	0
1	0	0	1	0	1	1	0
0	1	0	1	0	1	1	0
0	1	0	1	0	1	1	0
0	1	0	1	0	1	1	0

Выбор модели

	Decision Tree Classifier	Random Forest Classifier	XGB Classifier	Logistic Regression	Ada Boost Classifier
f1 for e-mail	0.87	0.84	0.82	0.57	0.77
auc-pr for e-mail	0.79	0.94	0.92	0.67	0.86
gini for e-mail	0.81	0.94	0.90	0.64	0.83
f1 for sms	0.74	0.38	0.81	0.01	0.35
auc-pr for sms	0.57	0.70	0.90	0.22	0.66
gini for sms	0.70	0.87	0.97	0.47	0.88

Выбор модели

- Для **e-mail** лучше всего подойдет:
  1. **Random Forest**, если важно иметь наименьший процент игнорируемых рассылок
  2. **Decision Tree**, если важнее найти больше откликов на коммуникацию

- Для **sms** лучше всего подойдет **XGB Classifier**

Тип коммуникации	Потенциальное число откликов (тыс)	Выявленное число откликов (тыс)	Число игнорируемых рассылок (тыс)
e-mail	42,1	32,4 (1) / 36,6 (2)	3,3 (1) / 5,4 (2)
sms	14,6	10,9	1,7

Выводы и итоги

Для разных коммуникаций лучше применять разные модели

В зависимости от ограничений одна модель может быть выгоднее другой