Анализ отклика клиентов банка на коммуникацию

Выполнили:
Земцова Анастасия
Иванова Анастасия
Катунцев Валентин
Черных Анна

Руководитель:
Титова Наталия

ПЛАН РАБОТЫ

ВВОДНАЯ ЧАСТЬ
- постановка целей
- описание данных

ИССЛЕДОВАНИЕ ДАННЫХ
- оценка пустых значений
- распределение данных
- поиск ошибочных значений
- подсчет уникальных значений

КОДИРОВАНИЕ ДАННЫХ
- подготовка витрины для построение моделей

ПОСТРОЕНИЕ МОДЕЛЕЙ
- сравнение разных моделей
- выбор наиболее подходящей для поставленной цели

ЗАКЛЮЧЕНИЕ
- финансово-экономические результаты кампании
- выводы

ВВОДНАЯ ЧАСТЬ: ЦЕЛЬ, ОПИСАНИЕ ДАННЫХ

Цель: построение аналитической модели, предсказывающую факт отклика клиента на коммуникацию

Показатель качества: статистика Accuracy

Представлена:
информация об анкетных данных клиента, его транзакционной активности в банке и история по откликам на коммуникации

Данные состоят из:

- 22 переменных: 13 категориальных, 9 количественных
- 985 477 строк (ID)

Name	Type	Label
ID	Character	ID клиента
Ind_Household	Character	Факт домовладения
Age_group	Character	Возрастная группа
District	Character	Район
Region	Character	Регион
Segment	Character	Статус клиента
Ind_deposit	Character	Индикатор владения депозитом
Ind_email	Character	Индикатор наличия e-mail
Ind_phone	Character	Индикатор наличия телефона
Ind_salary	Character	Индикатор владения зарплатной картой
Gender	Character	Пол
Target1	Character	Отклик на коммуникацию по e-mail
Target2	Character	Отклик на коммуникацию по телефону
Age	Numeric	Возраст
Lifetime	Numeric	Время, проведенное с банком
Income	Numeric	Доход
trans_6_month	Numeric	Транзакции за 6 месяцев
trans_9_month	Numeric	Транзакции за 9 месяцев
trans_12_month	Numeric	Транзакции за 12 месяцев
amont_trans	Numeric	Кол-во транзакций
amont_day_from	Numeric	Количество дней с последней транзакции
trans_3_month	Numeric	Транзакции за 3 месяца

ОБРАБОТКА ДАННЫХ: поиск пустых и ошибочных значений

Пустые значения были обнаружены в:

- Age = 66 958 (6,8%)
- Lifetime = 12 608 (1,3%)

Отсутствующие значения в столбце Age можно **заполнить средним значением** (54 года). После этого значения Unknown в столбце Age_group можно заменить соответствующей возрастной группой.

```
df_train['Age'] = df_train['Age'].fillna(float('54'))
```

```
df_train["Age_group"].replace("unknown", "middle", inplace=True)
```

Клиентов больше всего из South East

Ячейки со значением Unknown в столбце Region заменим значением South East

ИТОГ:

```
df_train.isna().sum()

ID                     0
Age                    0
Ind_Household          0
Age_group              0
District               0
Region                 0
Lifetime           12608
Income                 0
Segment                0
Ind_deposit            0
Ind_email              0
Ind_phone              0
Ind_salary             0
trans_6_month          0
trans_9_month          0
trans_12_month         0
amont_trans            0
amont_day_from         0
trans_3_month          0
Gender                 0
Target1                0
Target2                0
dtype: int64
```

1,3% от общего количества значений

Ячейки с неправильными значениями в:

- Age_group = 66 958 (6,85%)
- District = 29 752 (3%)
- Region = 20 510 (2,1%)
- Gender = 189 123 (19,2%)

Заменяем неправильные значение средними

ОБРАБОТКА ДАННЫХ: количество уникальных значений

```
count     985477
unique         4
top       middle
freq      578936
Name: Age_group, dtype: object
```

```
count     985477
unique        56
top           52
freq       53206
Name: District, dtype: object
```

```
count     985477
unique         2
top          Yes
freq      961490
Name: Ind_phone, dtype: object
```

```
count          985477
unique              6
top        South East
freq           382905
Name: Region, dtype: object
```

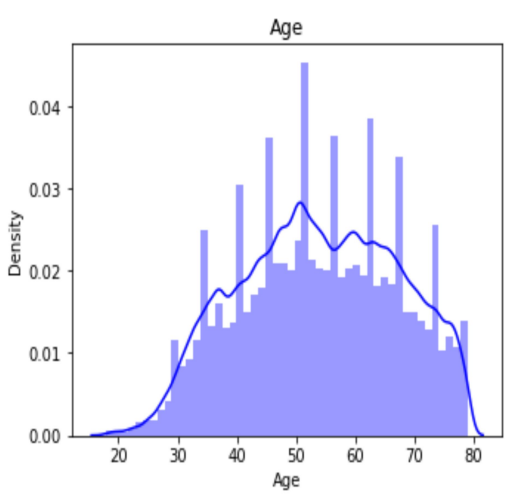
```
count     985477
unique         4
top       Silver
freq      379739
Name: Segment, dtype: object
```

```
count     985477
unique         2
top           No
freq      916207
Name: Ind_salary, dtype: object
```

Портрет клиента складывается из следующих показателей:

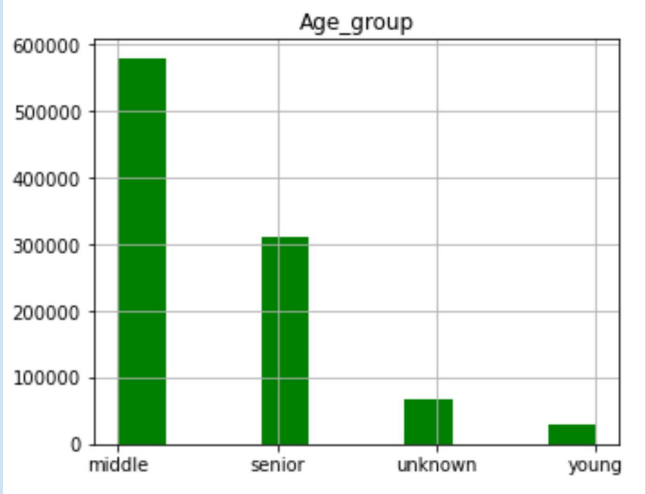
- возрастная группа (senior, middle, young)
- 6 регионов проживания и 56 райнов
- наличие депозита
- статус клиента в компании (Platinum, Gold, Silver, Tin)
- владение зарплатной картой (да/нет)

ИССЛЕДОВАНИЕ ДАННЫХ: Распределение данных



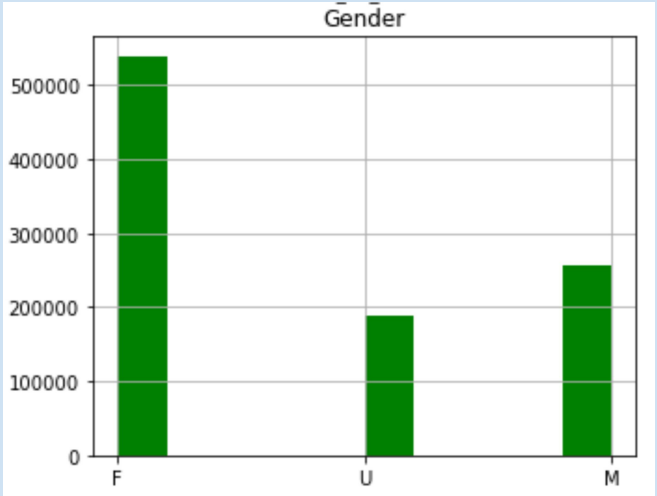
Средний возраст клиента:
54 года

```
df_train['Age'].describe()

count    918519.000000
mean         53.792106
std          13.196993
min          18.000000
25%          44.000000
50%          54.000000
75%          64.000000
max          79.000000
```



Больше всего откликов было от клиентов **среднего** возраста. Меньше всего откликались молодые клиенты

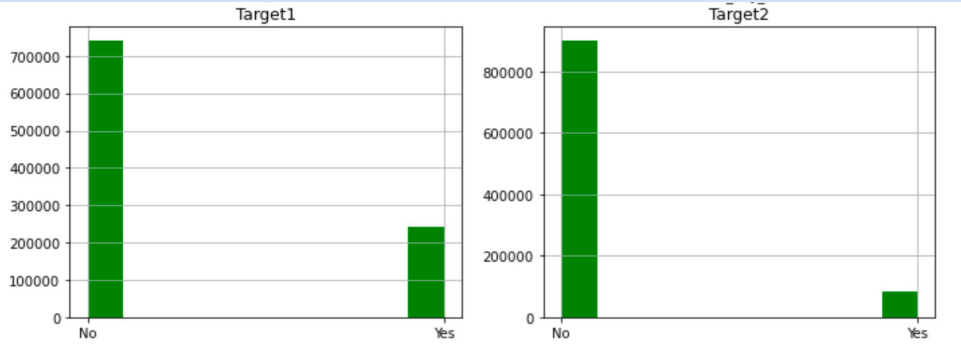
```
print(middle['Target1'].value_counts())
print(middle['Target2'].value_counts())

No      416862
Yes     162074
Name: Target1, dtype: int64
No      522840
Yes      56096
Name: Target2, dtype: int64
```



Больше половины клиентов **женского пола** (538 741 чел.: 54%) По 189 123 клиентам отсутствуют данные о поле, что составляет 19% от общего числа клиентов.

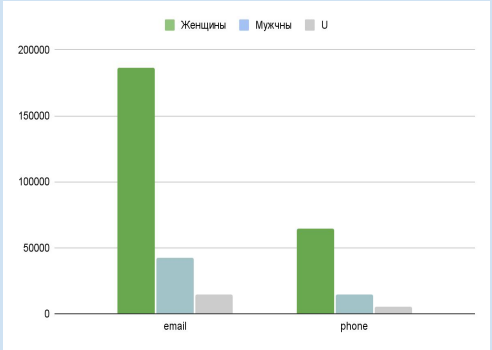
```
F = 538741
U = 189123
M = 257613
```

ИССЛЕДОВАНИЕ ДАННЫХ: Распределение данных

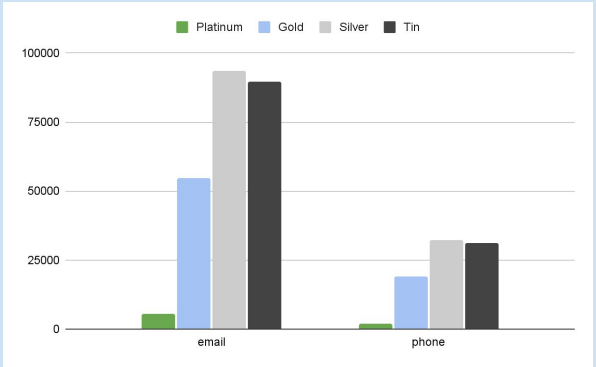


243 506 (25%) клиентов откликнулись на email

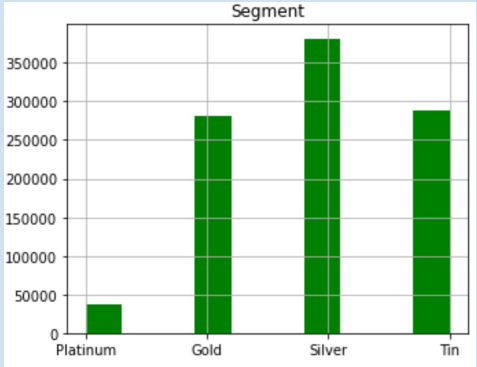
На телефон откликнулись **84 384 (8.5%)** клиента



Отклики на email и телефон в зависимости от пола



Отклики на email и телефон в зависимости от статуса



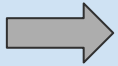
Распределение клиентов банка на сегменты

КОДИРОВАНИЕ ДАННЫХ:

Заменяем значение YES и NO на 1 и 0 для:
Target 1 и Target 2

	Target1	Target2
0	No	No
1	No	No
2	No	No
3	No	No
4	Yes	No
...	...	...
985472	No	No
985473	No	No
985474	No	No
985475	No	No
985476	No	No

[985477 rows x 2 columns]



	Target1	Target2
0	1.0	1.0
1	1.0	1.0
2	1.0	1.0
3	1.0	1.0
4	0.0	1.0
...	...	...
985472	1.0	1.0
985473	1.0	1.0
985474	1.0	1.0
985475	1.0	1.0
985476	1.0	1.0

[985477 rows x 2 columns]

```
def one_hotizer(df, columns):
    new_df = df.copy()
    for col in columns:
        new_df[col] = encode(new_df[col])
    return new_df
```

Убираем столбец 'Age_group', т. к. столбец 'Age' уже содержит достаточно информации о возрасте клиентов

```
df_train.drop(labels='Age_group', axis=1, inplace=True)
```

Модели

Были рассмотрены модели:

- **Дерево решений** (DecisionTree)
- **Случайный лес** (RandomForestClassifier)
- **Логистическая регрессия** (LogisticRegression)
- **Градиентный бустинг (**XGBoost)

Для подбора параметров модели использовались:

- **RandomizedSearchCV**
- **GridSearchCV**

```
from sklearn.model_selection import RandomizedSearchCV
from sklearn.model_selection import GridSearchCV
```

Метрики качества:

- **AUC-ROC**
- **f1 = (2*precision*recall)/(precision+recall)**

СРАВНЕНИЕ МОДЕЛЕЙ:

Best model

		Decision tree	Logistic regression	Random Forest	XGBoost
TARGET 1 (отклик по email)	f1 test	0.764572	0.548718	0.714025	0.773504
	f1 train	0.763685	0.544586	0.712395	0.77284
	AUC-ROC test	0.924629	0.829914	0.924185	0.931258
	AUC-ROC train	0.924593	0.827963	0.925471	0.931107
TARGET 1 (отклик по phone)	f1 test	0.441479	0.005767	0.0	0.744886
	f1 train	0.439915	0.006167	0.000067	0.746809
	AUC-ROC test	0.92256	0.755483	0.873042	0.966241
	AUC-ROC train	0.922415	0.754065	0.879054	0.966604

ФИНАНСОВО-ЭКОНОМИЧЕСКИЕ РЕЗУЛЬТАТЫ КАМПАНИИ

канал отклика	ед измерения	email			телефон			итого
возраст		young	middle	senior	young	middle	senior	
Response (Отклик)		14%	7%	4%	2%	1%	0%	
Circulation (Объем рассылки)	ед.	325	6400	3435	936	18500	9900	
доход от клиентов		2 909 ₽	2 909 ₽	2 909 ₽	2 909 ₽	2 909 ₽	2 909 ₽	
Выручка		134 816 ₽	1 287 752 ₽	371 737 ₽	47 544 ₽	446 471 ₽	127 606 ₽	
Ограничение на бюджет		27 722 ₽	266 696 ₽	76 905 ₽	9 807 ₽	92 409 ₽	26 461 ₽	500 000 ₽
Стоимость СМС					4 680 ₽	92 500 ₽	49 500 ₽	
Стоимость email								50 000 ₽
Затраты на привлечение клиентов(за исключением затрат)		27 809 ₽	265 630 ₽	76 680 ₽	9 807 ₽	92 095 ₽	26 322 ₽	498 343 ₽
Прибыль		57 007 ₽	972 122 ₽	245 057 ₽	33 057 ₽	261 875 ₽	51 784 ₽	1 620 903 ₽

Общий доход для банка составит 1.6 милл., при затратах 489 тыс.руб

Средняя сумма потребительского кредита: 268 т. при ставке 13% = 2909 рублей

Валовая доходность проекта составит 31%

ПОЛУЧЕННЫЕ РЕЗУЛЬТАТЫ

Цель: построение аналитической модели, предсказывающую факт отклика клиента на коммуникацию

Команда провела:

ИССЛЕДОВАНИЕ ДАННЫХ

- оценка пустых значений
- распределение данных
- поиск ошибочных значений
- подсчет уникальных значений

КОДИРОВАНИЕ ДАННЫХ

- подготовка витрины для построение моделей

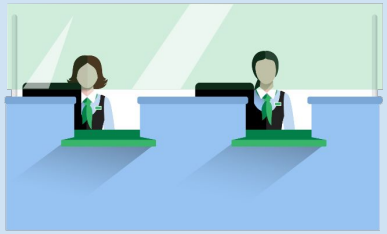
ПОСТРОЕНИЕ МОДЕЛЕЙ

- сравнение разных моделей
- выбор наиболее подходящей для поставленной цели (модель **XGBoost)**

ЗАКЛЮЧЕНИЕ

- финансово-экономические результаты кампании (общая сумма прибыли = 1 620 903)

СПАСИБО ЗА ВНИМАНИЕ



Github проекта:
https://github.com/ValentineKatuntsev/project.git