Прогноз відтоку абонентів

Євген Шнурко
Big Data Lab 5
Vodafone

Відтік абонентів

- Зменшення доходу
- Підвищення витрат
- Погіршення репутації
- Втрата конкурентних переваг

Зазвичай, залучення нових клієнтів дорожче, ніж утримання наявних.

Відтік 1000 абонентів це:

1,440,000 ₴ на рік

Зменшення прибутку *

300,000 ₴

Витрат на залучення такої ж кількості нових клієнтів**

* Виходячи з середнього чеку 120 ₴ на місяць

** Виходячи з вартості залучення нового абоненту 300 ₴

Мета проекту:

- Ідентифікація клієнтів, які можуть залишити компанію.

Задля:

- Своєчасних та ефективних заходів для утримання. Зокрема, через спеціальні пропозиції, персоналізовані кампанії, покращення обслуговування, тощо.

1000 абонентів, що заздалегідь розпізнані як ті, що йдуть у відтік, це:

$$(1{,}440{,}000 \cdot K)\ ₴ \text{ на рік}$$

На які ми зменшуємо втрату доходу.

Де K – конверсія, міра ефективності комунікації.

$0 \leq K \leq 1$

Дані:

- Основний датасет: 3хмісячний зріз активності абонентів, 150 000 спостережень, 817 ознак.
- Додатковий датасет B_NUM: дані про короткі номери, які взаємодіяли з абонентами, 671 248 спостережень, 8 ознак.
- Додатковий датасет DPI: трафік мобільних додатків абонентів, 6 745 887 спостережень, 6 ознак.

Розподіл значень цільової змінної:

Навчальний набір даних:		**Тестовий набір даних:**	
0	140 414	0	140597
1	9 586	1	9403

Всі ознаки приймають виключно числові значення.

201 ознака має пропущені значення.

Неінформативні ознаки:

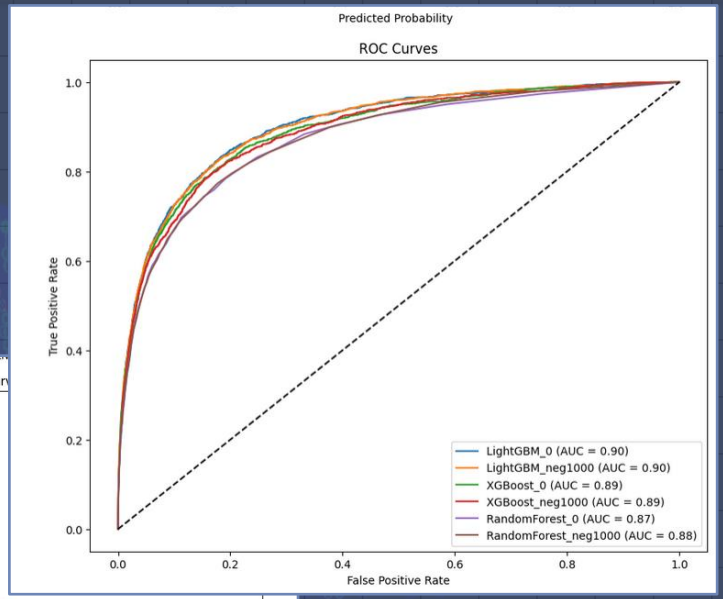
- 11 ознак, що не приймають жодного значення;
- 33 ознаки, що мають лише одне унікальне значення та не мають пропущених значень;
- 2 ознаки що мають відповідно 2 та 6 непорожних спостереження, що розподілені по обох класах;
- Ідентифікатор абонента.

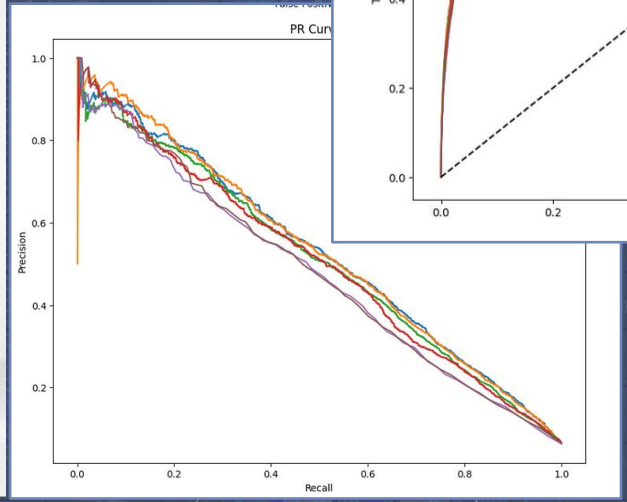
Моделі:

- LightGBM (Light Gradient Boosting Machine)
- XGBoost (Extreme Gradient Boosting)
- Random Forest

Тренування та валідація проводилися на розбитому у співвідношенні 80 до 20 навчальному датасеті. Фінальне тестування роботи моделі проводилося на тестовому датасеті.

Моделі:





AMD Ryzen 5 5500U with Radeon Graphics, 2.10 GHz

	LightGBM	XGBoost	Random Forest
Training Time, sec:	18.36	37.20	386.63
Prediction Time , sec :	3.80	2.49	25.96

Первинні результати:

- Залишаємо пропущені значення як є;
- Найкращі результати модель демонструє на топ 25, 241 та 454 найважливіших ознак.

Балансування класів збільшує охоплення цільової групи, але робить нагальним питання помилково розпізнаних як ті, що йдуть у відтік.

Додаткові датасети:

Датасет BNUM:

- 2301 додаткова ознака, що характерізує взаємодію абонента з кожним з коротких номерів.

Датасет DPI:

- 5 додаткових ознак, що характерізують взаємодію абонента із застосунками взагалі;
- 2976 додаткових ознак, що характерізують взаємодію абонента з кожним застосунком, по 744 ознаки на кожну з 4 метрик.

Найкращі результати:

- 25 найкращих ознак основного датасету та 744 ознаки з датасету DPI;
- Залишено 40 найважливіших, до яких увійшли 25 ознак основного датасету та 15 згенерованих;
- Для знайденої комбінації ознак були підібрані оптимальні параметри моделі.

```
params = {
    'boosting_type': 'gbdt',
    'num_leaves': 24,
    'max_depth': 6,
    'learning_rate': 0.010606
    'n_estimators': 985,
    'reg_alpha': 9.4206459046
    'reg_lambda': 0.101262332
    'min_split_gain': 0.09655
    'subsample': 0.9931456581
    'colsample_bytree': 0.766
    'objective': 'binary',
    'metric': 'auc',
    'is_unbalance': True
}
```

Результати моделі:

TEST:

AUC: 0.8971

Recall: 0.7162

FP/TP Ratio: 1.9868

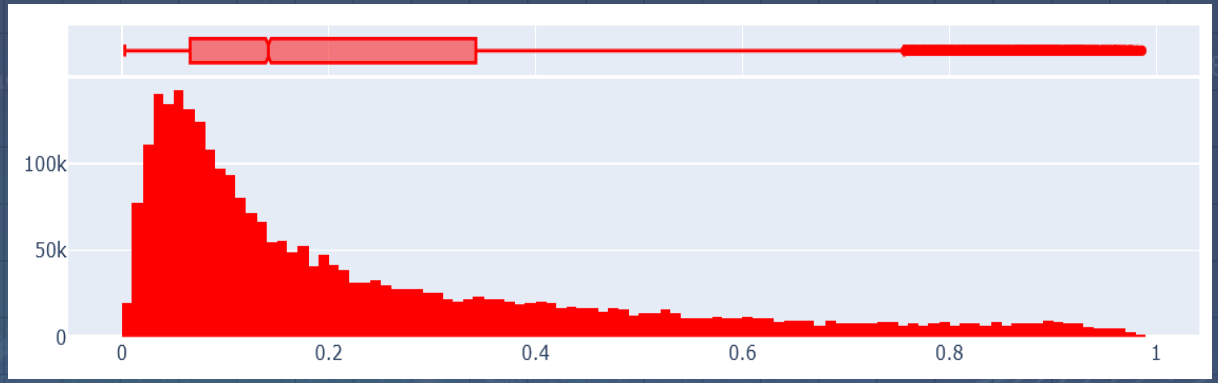
TRAIN:

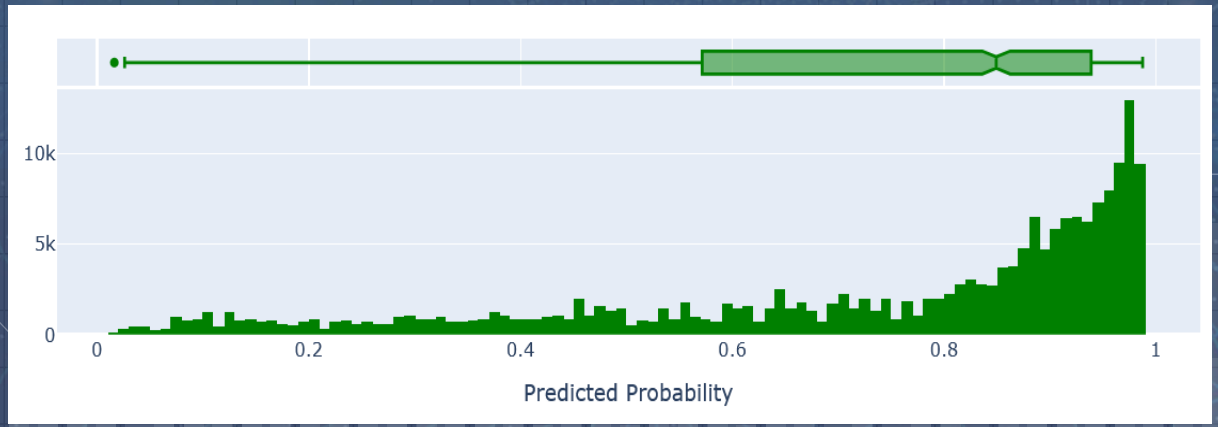
AUC Train: 0.9302

AUC Validation: 0.9019

Recall: 0.7136

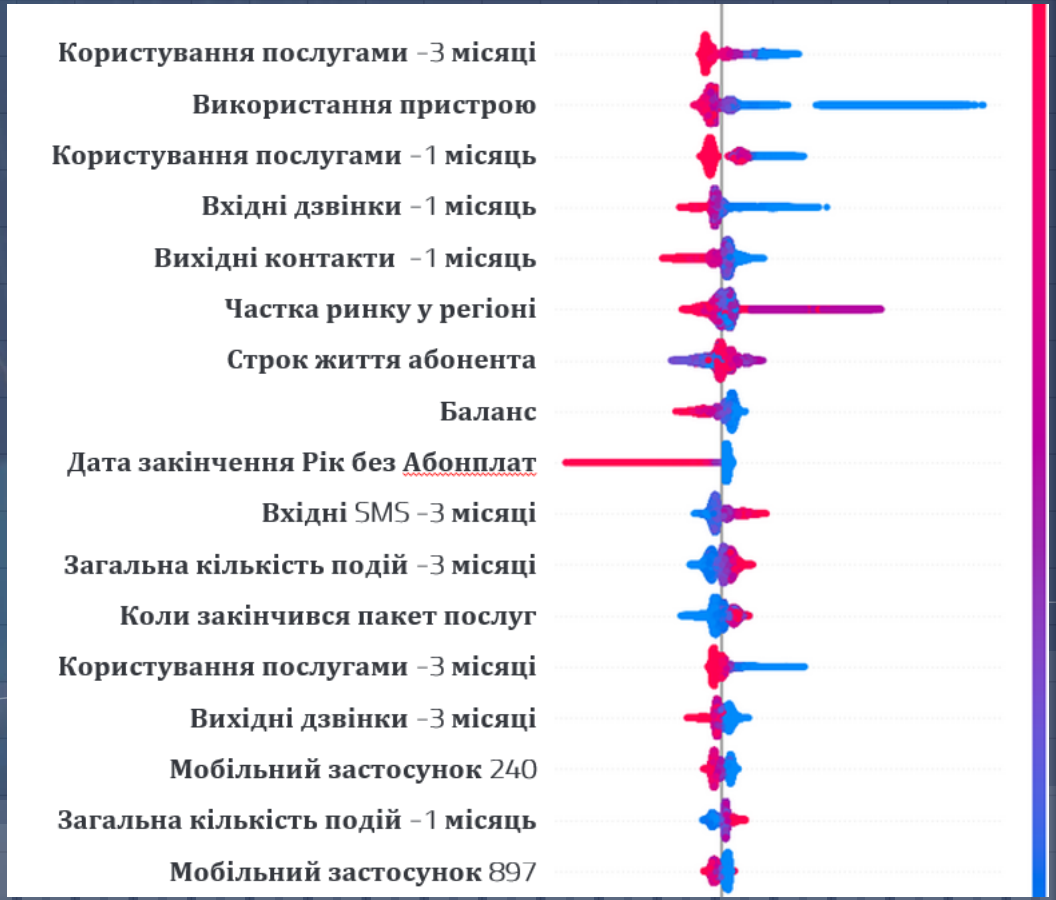
FP/TP Ratio: 1.9094



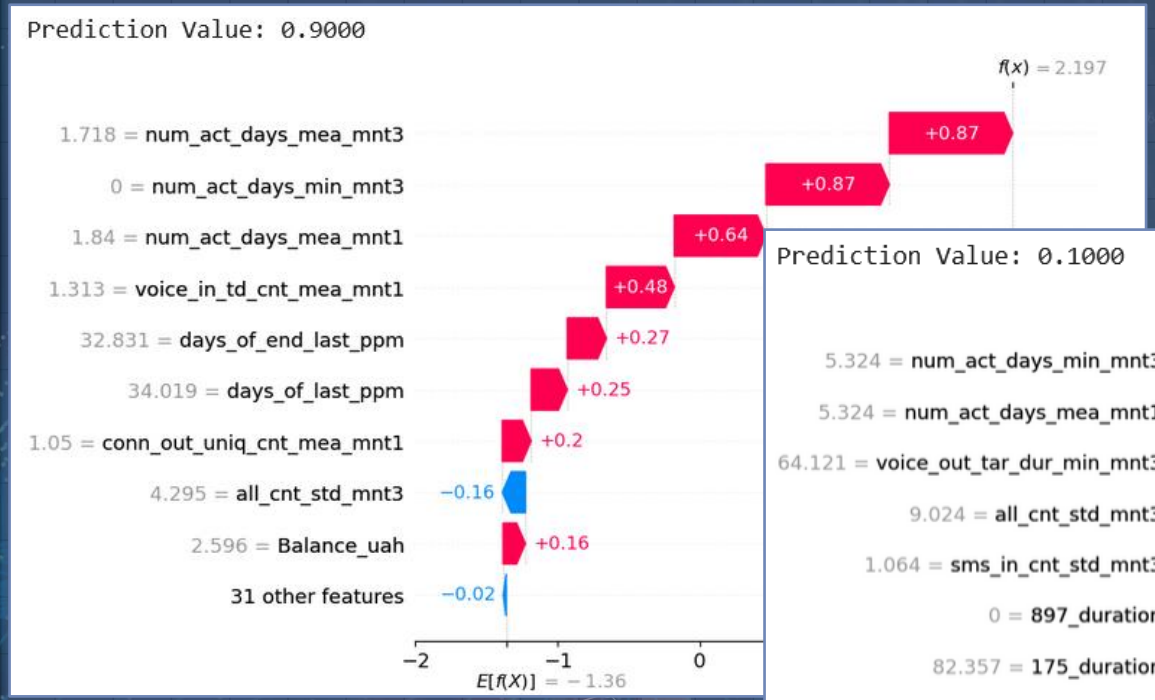


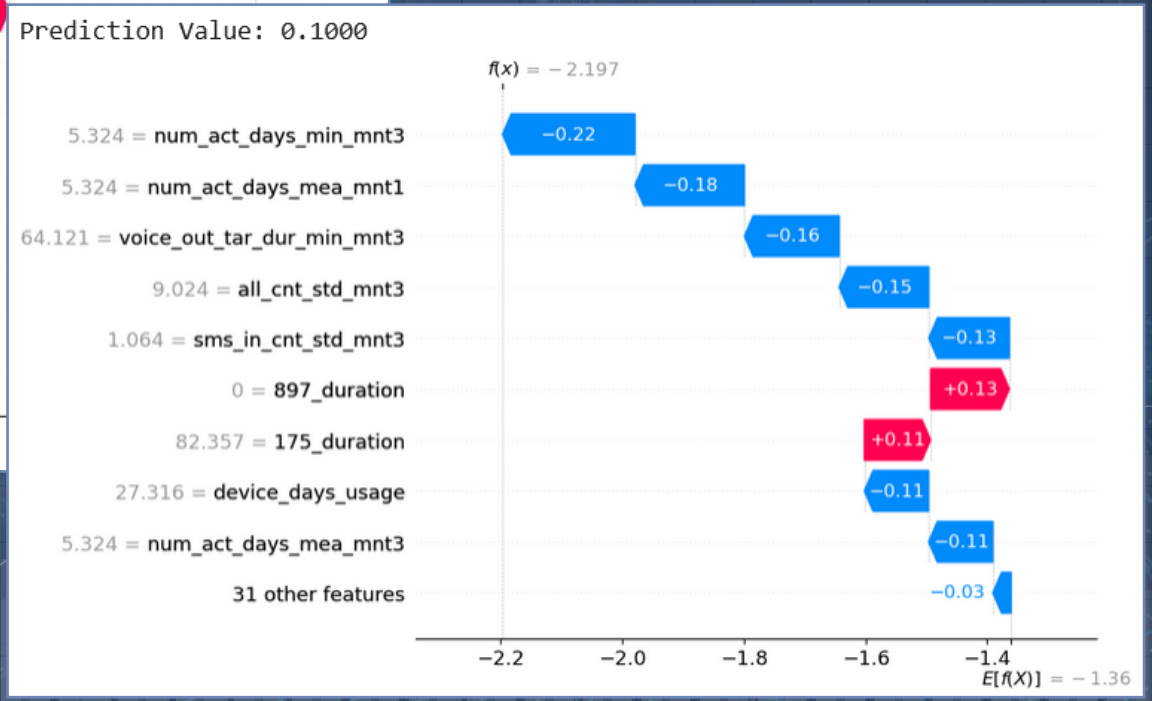
Інтерпретація результатів

- Зменшення споживання послуг
- Використання пристрою
- Певні мобільні застосунки

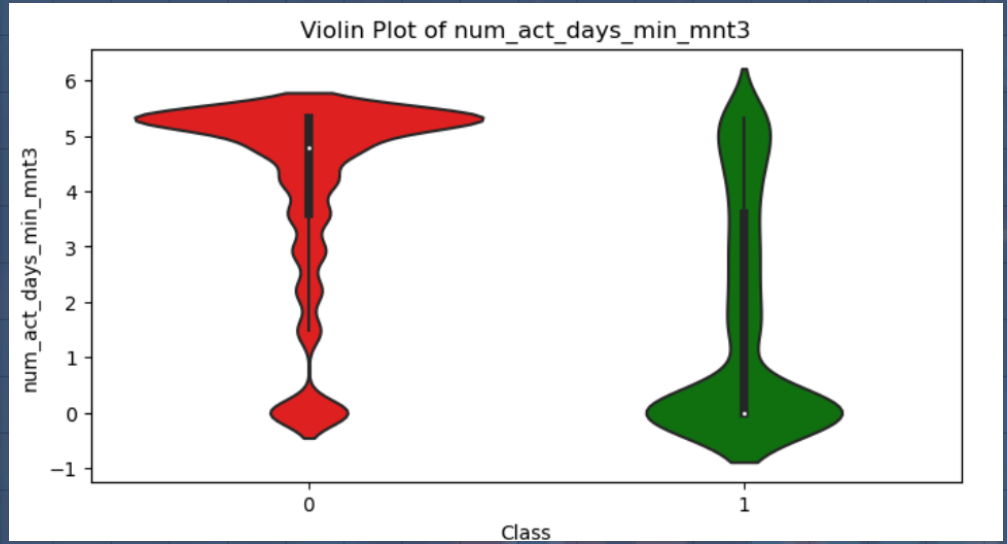


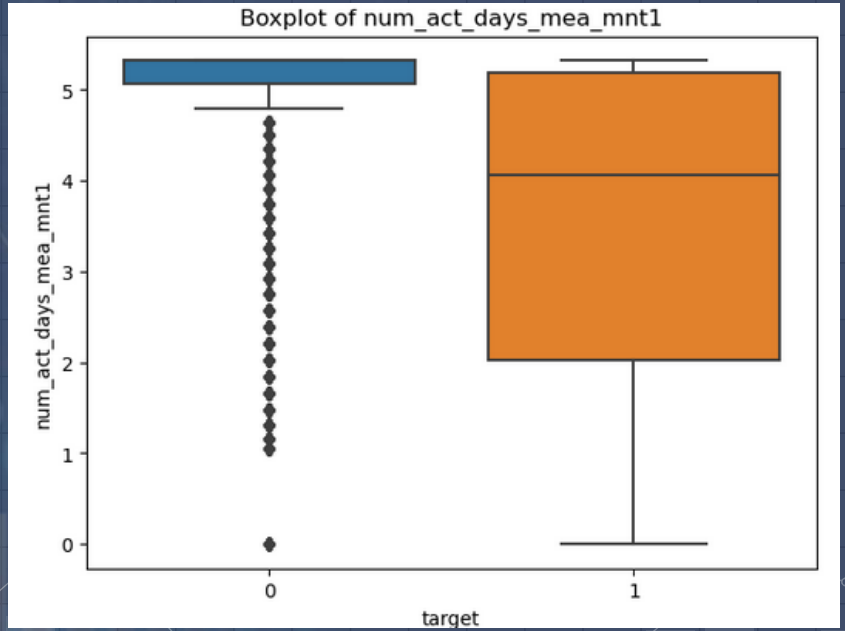
Інтерпретація результатів





Зменшення споживання





Спостерігається як у ознаках що стосуються періоду –3 місяці, так і –1 місяць.

Recall vs False Positives:

Скільки помилково розпізнаних як такі, що йдуть у відтік ми можемо собі дозволити?

Recall vs False Positives:

Скільки помилково розпізнаних як такі, що йдуть у відтік ми можемо собі дозволити?

Наприклад, якщо ми пропонуємо знижку 50%:

Recall vs False Positives:

Скільки помилково розпізнаних як такі, що йдуть у відтік ми можемо собі дозволити?

Наприклад, якщо ми пропонуємо знижку 50%:

0,5 · **120₴** · **12** місяців · К-сть залишившихся

Recall vs False Positives:

Скільки помилково розпізнаних як такі, що йдуть у відтік ми можемо собі дозволити?

Наприклад, якщо ми пропонуємо знижку 50%:

0,5 · **120**₴ · **12** місяців · Вірно розпізнані · K_1 =

Recall vs False Positives:

Скільки помилково розпізнаних як такі, що йдуть у відтік ми можемо собі дозволити?

Наприклад, якщо ми пропонуємо знижку 50%:

$$0{,}5 \cdot 120₴ \cdot 12 \text{ місяців} \cdot \text{Вірно розпізнані} \cdot K_1 =$$
$$= 0{,}5 \cdot 120₴ \cdot 12 \text{ місяців} \cdot \text{Помилково роспізнані} \cdot K_0$$

Recall vs False Positives:

Скільки помилково розпізнаних як такі, що йдуть у відтік ми можемо собі дозволити?

Наприклад, якщо ми пропонуємо знижку 50%:

$$= 0{,}5 \cdot 120€ \cdot 12 \text{ місяців} \cdot \text{Вірно розпізнані} \cdot K_1 =$$
$$= 0{,}5 \cdot 120€ \cdot 12 \text{ місяців} \cdot \text{Помилково розпізнані} \cdot K_0$$

Recall vs False Positives:

Скільки помилково розпізнаних як такі, що йдуть у відтік ми можемо собі дозволити?

Наприклад, якщо ми пропонуємо знижку 50%:

$$\frac{\text{Помилково розпізнані}}{\text{Вірно розпізнані}} = \frac{K_1}{K_0}$$

Recall vs False Positives:

Скільки помилково розпізнаних як такі, що йдуть у відтік ми можемо собі дозволити?

Або у більш загальному випадку:

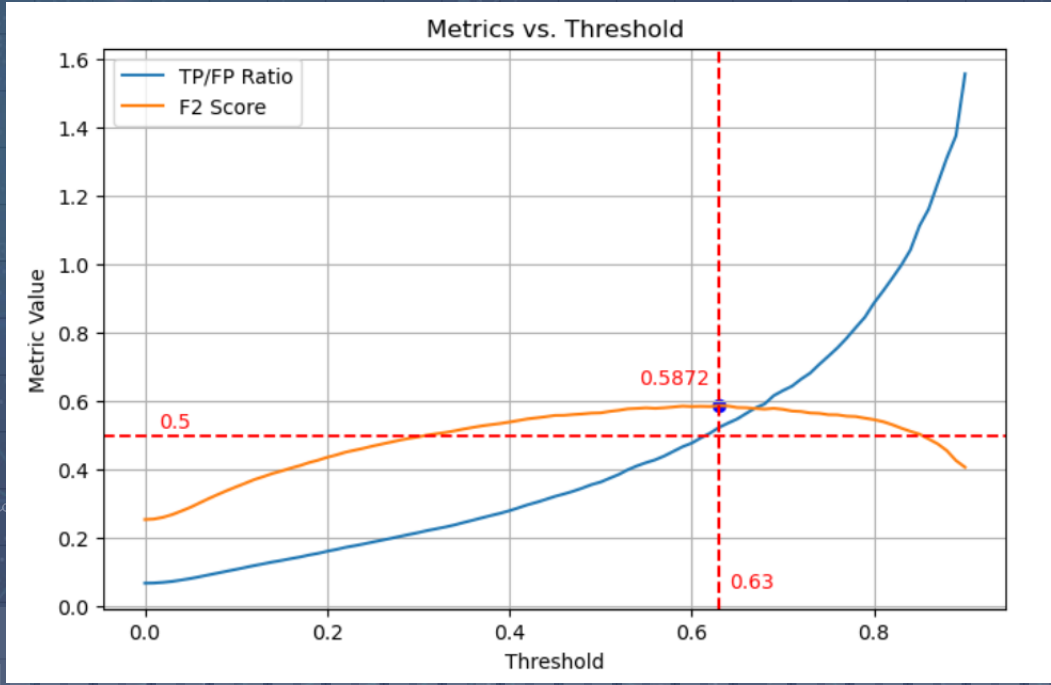
$$\frac{\text{Помилково розпізнані}}{\text{Вірно розпізнані}} = \frac{K_1}{K_0} \cdot \frac{\text{Знижка,\%}}{(1 - \text{Знижка,\%})} = R$$

До співвідношенням конверсій додається співвідношення прибутку та витрат.

Recall vs False Positives:

Потрібне співвідношення досягається за рахунок зміни порогу.

Максимум
$F_2 = 0{,}59$
що відповідає
$R = 2\ (1/\ R = 0{,}5)$
при порозі 0,63



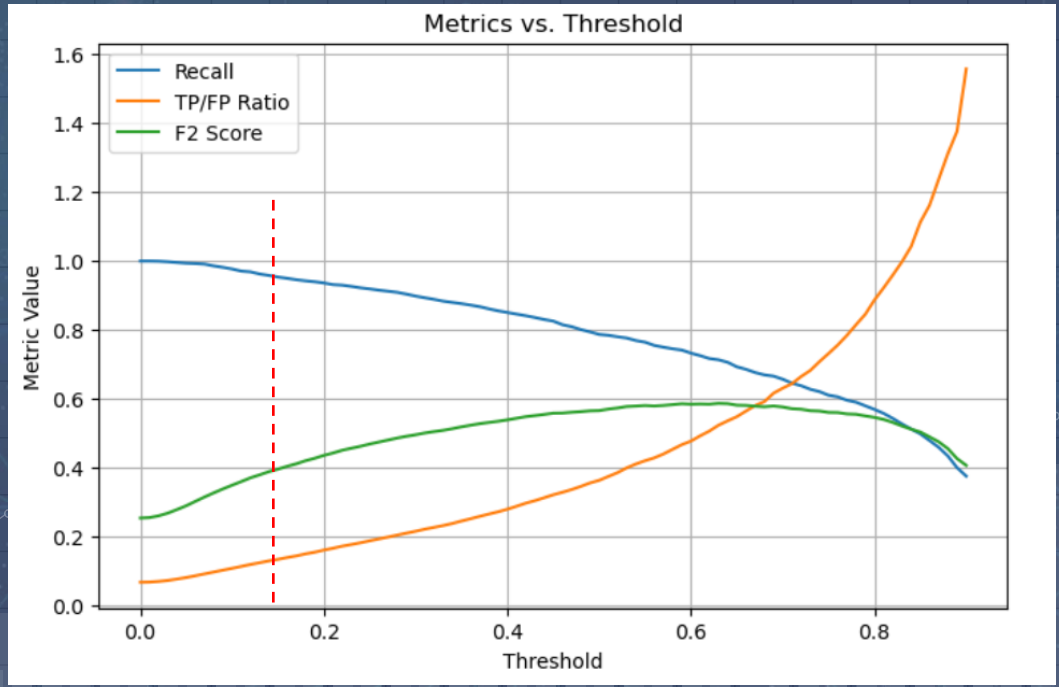
Recall vs False Positives:

При вдвічі кращій конверсії у цільовій групі та співвідношенні прибутку до витрат як 3 до 1:

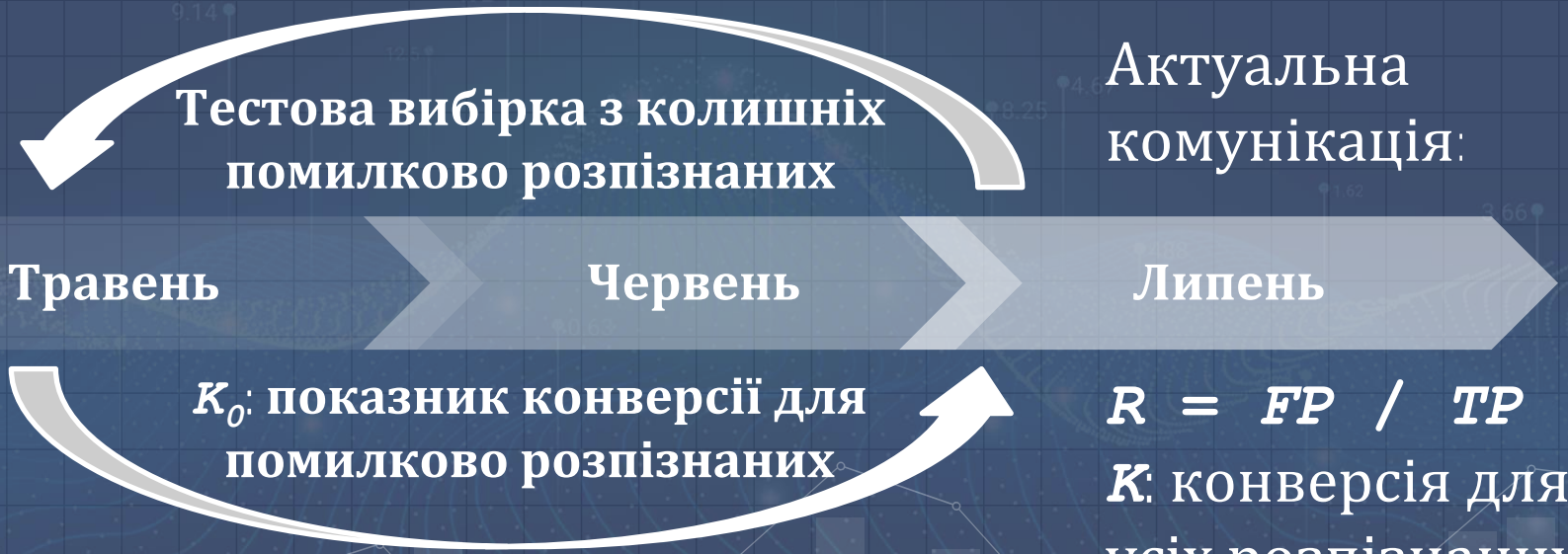
$R = 6$

$F_2 = 0{,}4$

Recall = 0.93



Обчислення конверсій



Актуальна комунікація:

$R = FP / TP$

K: конверсія для усіх розпізнаних

K_1 – ?

Обчислення конверсій

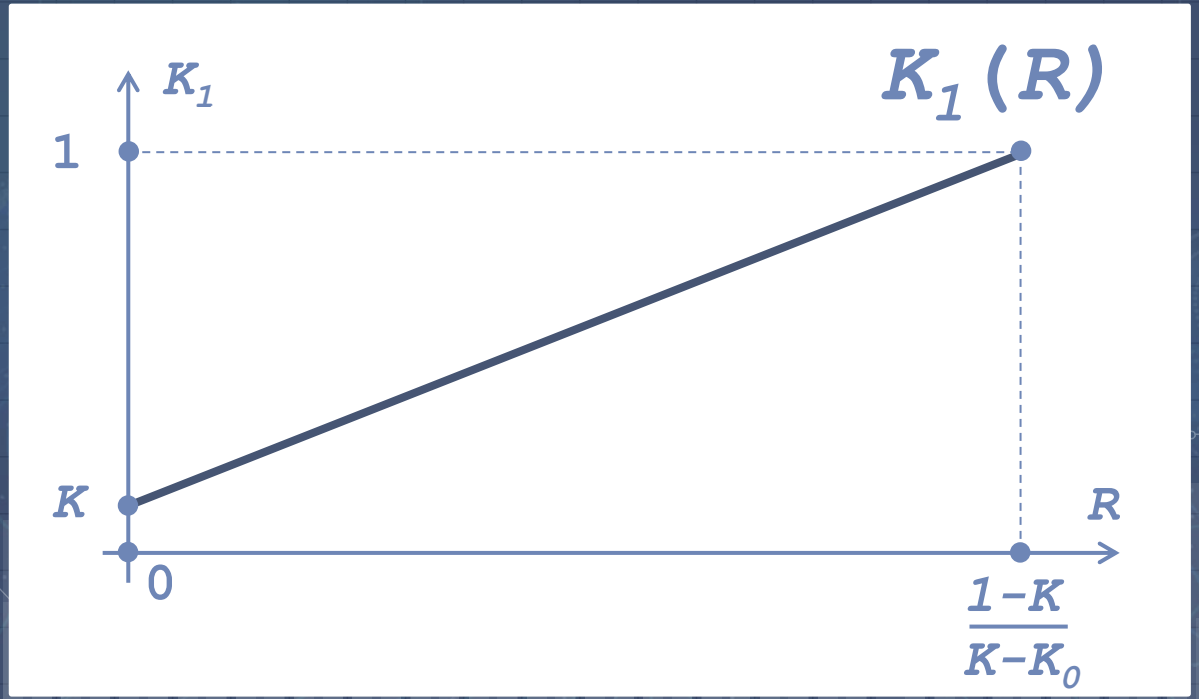
$$K_1 = R \cdot (K - K_0) + K$$

$K_1 > K$

$K > K_0$

$R = FP/TP$

$$K = \frac{FP \cdot K_0 + TP \cdot K_1}{FP + TP}$$

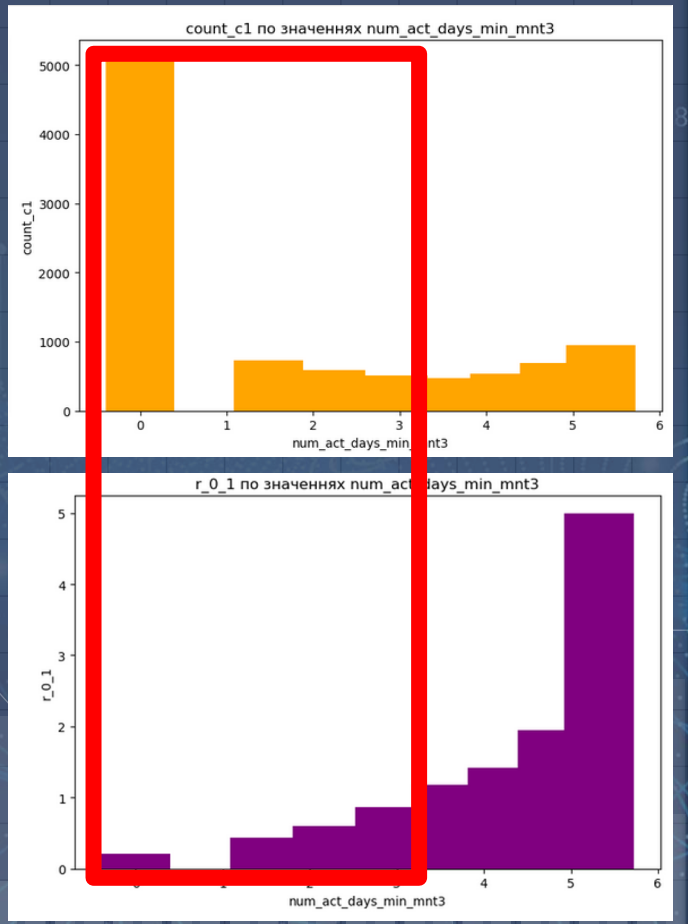


Кластерізація даних

Кількість днів активного користування послугами -3 місяці

Охоплення цільової групи: 72%
(для 4 значень ознаки з 8)

AUC Test: 0.83
Recall Test: 0.6930
FP/TP Ratio Test: 0.84

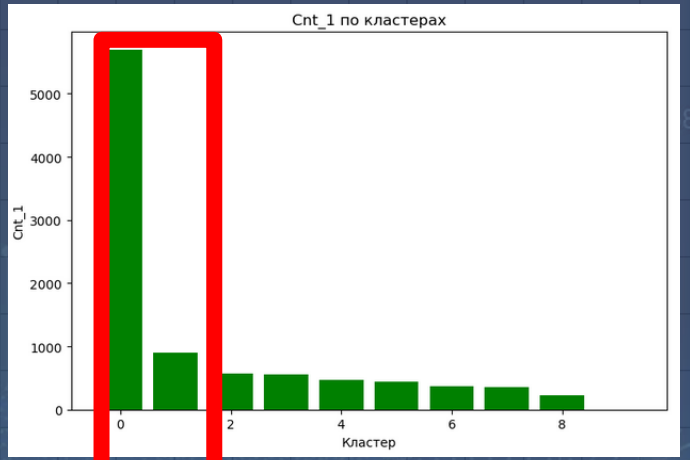


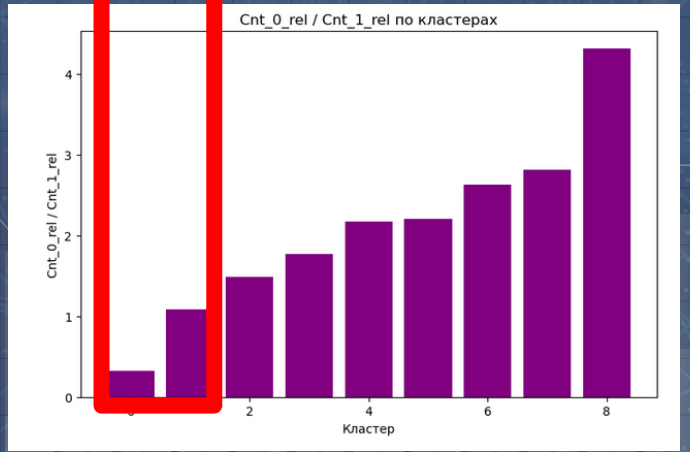
Кластерізація даних

Кількість унікальних вихідних контактів -1 місяць

Охоплення цільової групи: 68%
(для 2 кластерів з 10)
AUC Test: 0.86
Recall Test: 0.72
FP/TP Ratio Test: 1.47

Охоплення цільової групи: 59%
(для 1 кластера з 10)
AUC Test: 0.85
Recall Test: 0.70
FP/TP Ratio Test: 1.21





Кластерізація даних

Вхідні SMS -3 місяці: Охоплення 51% (2 кластери з 10)

AUC Test: 0.89 Recall Test: 0.76 FP/TP Ratio Test: 1.60

Користування послугами (сер.): Охоплення 64% (2 кластери з 10)

AUC Test: 0.83 Recall Test: 0.69 FP/TP Ratio Test: 1.07

Кількість подій –3 місяці: Охоплення 51% (4 кластери з 10)

AUC Test: 0.90 Recall Test: 0.76 FP/TP Ratio Test: 1.71

Висновки

- Зменшення споживання послуг потребує негайних дій;
- Комунікації повинні максимізувати конверсію у цільовій групі та мінімізувати конверсію для помилково розпізнаних;
- Використання мобільних застосунків;
- Жодна з ознак що стосується технічних проблем не увійшла до топ200 ознак за важливістю;
- Активність конкурентів вимагає протидії;
- Важливість частки ринку?

Дякую за увагу! Запитання?








shnurko@gmail.com
+380 67 405 24 78

Зображення зґенеровані за запитом "LightGBM predicts abonents churn for Vodafone"