**НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ**

**«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ  
імені ІГОРЯ СІКОРСЬКОГО»**

**Інститут прикладного системного аналізу**

**Кафедра математичних методів системного аналізу**

|  |  |
| --- | --- |
| «На правах рукопису»  УДК 336.77:330.131.7 | «До захисту допущено»  Завідувач кафедри  \_\_\_\_\_\_\_\_\_\_ О.Л. Тимощук  «\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_\_20\_\_ р. |

**Магістерська дисертація**

**на здобуття ступеня магістра**

**зі спеціальності 124 Системний аналіз**

**на тему: «Скорингові моделі для оцінювання платоспроможності абонентів телекомунікаційного оператора »**

Виконав:

студент ІІ курсу, групи КА-62м

Роговий Андрій Владиславович \_\_\_\_\_\_\_\_\_\_

Керівник:

д.т.н., проф.

Бідюк П.І. \_\_\_\_\_\_\_\_\_\_

Рецензент:

д.т.н., проф.

Архипов О.Е. \_\_\_\_\_\_\_\_\_\_

Засвідчую, що у цій магістерській дисертації немає запозичень з праць інших авторів без відповідних посилань.

Студент \_\_\_\_\_\_\_\_\_\_\_\_\_

Київ

2018

**Національний технічний університет України**

**«Київський політехнічний інститут імені Ігоря Сікорського»**

**Інститут прикладного системного аналізу**

**Кафедра математичних методів системного аналізу**

Рівень вищої освіти – другий (магістерський)

Спеціальність (спеціалізація) – 124 «Системний аналіз» («Системний аналіз і управління»

ЗАТВЕРДЖУЮ

Завідувач кафедри

\_\_\_\_\_\_\_\_\_\_ О.Л. Тимощук

«\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_\_20\_\_ р.

**ЗАВДАННЯ**

**на магістерську дисертацію студенту**

**Роговий Андрій Владиславович**

1. Тема дисертації «Скорингові моделі для оцінювання платоспроможності абонентів телекомунікаційного оператора», науковий керівник дисертації Бідюк Петро Іванович профессор, д.т.н., затверджені наказом по університету від «\_\_\_»\_\_\_\_\_\_\_\_\_ 20\_\_ р. №\_\_\_\_\_

2. Термін подання студентом дисертації \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

3. Об’єкт дослідження статистичні дані телекомунікаційного оператора “Київстар”

4. Предмет дослідження Математичні моделі та методи обробки статистичних даних, критерії якості даних та адекватності моделей

5. Перелік завдань, які потрібно розробити 1. Зібрати необхідні статистичні дані і вибрати методи моделювання для оцінювання кредитоспроможності клієнтів. 2. Побудувати моделі для прогнозу платоспроможності абонентів “Київстар” за даними компанії “Київстар” 3. Побудувати моделі для прогнозу платоспроможності абонентів інших операторів. 4. Виконати аналіз отриманих результатів

6. Орієнтовний перелік графічного (ілюстративного) матеріалу 1. Актуальність теми методи та мета дослідження. 2. Постановка задачі. 3. Джерела кредитної інформації 4. Що можна дізнатися з даних телеком-оператора ? 5. Препроцессінг даних. 6. Logistic regression. 7. Decission tree. 8. Ансамблы дерев. 9. Побудова моделей (абоненти “Київстар”) 10. Результати роботи моделі (абоненти “Київстар”). 11. Побудова моделей (інші абоненти). 12. Результати роботи моделі (інші оператори). 13. Висновки.

7. Орієнтовний перелік публікацій: Тези конференції САІТ\_\_

8. Дата видачі завдання \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Календарний план

|  |  |  |  |
| --- | --- | --- | --- |
| № з/п | Назва етапів виконання  магістерської дисертації | Термін виконання етапів магістерської дисертації | Примітка |
| 1 | Формування теми дослідження | 09. 2016 |  |
| 2 | Збір статистичних даних | 05.2017 |  |
| 3 | Виконання обчислювальних експериментів та моделювання | 11. 2017 |  |
| 4 | Аналіз результатів обчислювальних експериментів | 01.2018 |  |
| 5 | Корегування результатів моделювання та прогнозування | 03. 2018 |  |
| 6 | Підготування тексту дисертаційного дослідження | 04. 2018 |  |
| 7. | Попередній захист | 04. 2018 |  |
| 8. | Надання роботи до захисту | 05. 2018 |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Студент А.В. Роговий

Науковий керівник дисертації П.І.Бідюк

РЕФЕРАТ

Магістрська дисертація: 99 с., 20 рисунків, 27 таблиць., 3 додатка, 23 джерел.

Актуальність теми – у банківській сфері при управлінні кредитними ризиками одним з ключових завдань є оцінка кредитоспроможності позичальників. Незважаючи на великий світовий досвід кредитування осіб, з кожним роком з’являється все більша необхідність в покращенні скорингових моделей, у зв’язку з цим банки постійно шукають нові не стандартні джерела кредитування. Кредитування за телеком-даними є новою задачею, яка потребує для розв`язання окрім необхідних знань, також генерування нових ідей та методів.

Об’єкт дослідження – дані телекомунікаційного оператора “Київстар” на основі яких виконувалося прогнозування кредитоспроможності клієнтів.

Мета роботи – розробка моделей для оцінювання платоспроможності абонентів “Київстар” та інших операторів за даними компанії “Київстар”.

Методи дослідження: логістична регресія, градієнтний бустинг, Random Forest.

Отримані результати: математичні моделі для оцінювання платоспроможності абонентів “Київстар” та інших операторів за даними компанії “Київстар”.

Публікації – публікація на науковій конференції “САІТ”.

Магістрська дисертація має акт впровадження.

Дослідження виконанні з використанням мов програмування Python та PL/SQL.

КРЕДИТНИЙ РИЗИК, ПРОГНОЗУВАННЯ, АНАЛІЗ СТАТИСТИЧНИХ ДАНИХ, МАТЕМАТИЧНА МОДЕЛЬ, ТЕЛЕКОМ-ОПЕРАТОР.

ABSTRACT

Master`s thesis: 99 p., 20 figures, 27 tables, 3 appendixes, 23 sources.

Actuality - in the banking sector in the management of credit risks one of the key tasks is to assess the creditworthiness of borrowers. Despite the great global experience of lending to individuals, every year there is an increasing need to improve scorecard models, in this connection, banks are constantly looking for new non-standard lending sources. Using the telecom data for the credit scoring is a rather new task, which requires, besides the necessary knowledge, the usage of new ideas and methods

Publications - publication at the scientific conference "SAIT"

The object of the research - the data of Kyivstar telecommunication operator on the basis of which the forecasting of the client's creditworthiness was performed.

Purpose - develop models for assessing the solvency of Kyivstar and other operators subscribers.

Research methods - logistic regression, Gradient Boosting, Random Forest.

Obtained results - a model for assessing the solvency of Kyivstar subscribers and other operators. Approaches to primary data processing were investigated and a comparative analysis of the results of several models was conducted.

Work has an act of implementation.

The research was conducted using Python and PL / SQL programming languages.

CREDIT RISK, FORECAST, ANALYSIS, MODEL, TELECOM OPERATOR, DATA SCIENCE.

ЗМІСТ

[**ПЕРЕЛІК ПРИЙНЯТИХ ПОЗНАЧЕНЬ ТА СКОРОЧЕНЬ**](#_ixsrypv4chpg) **9**

[**ВСТУП**](#_1v9tyqnrnvev) **10**

[**РОЗДІЛ 1 ОЦІНКА ПЛАТОСПРОМОЖНОСТІ КЛІЄНТА**](#_ldz9x9s4c5m2) **11**

[1.1 Сутність кредитного ризику і методи боротьби з ним](#_tz5og59i3q64) 11

[1.2 Нові джерела даних](#_x88qenhqsoj) 13

[1.2.1 Причини залучення нових даних](#_tig4r39jpy08) 13

[1.2.2 Дані соціальних мереж](#_aggaign5g9o) 14

[1.2.3 Дані “Яндекс Таксі” або Uber](#_uwkyqggmdc8a) 16

[1.2.4 Дані про поведінку користувача на сайті банку](#_ivut2lfif4b6) 17

[1.2.5 Дані телеком-оператора](#_fni0btbvwhbd) 18

[Висновки до розділу 1](#_3pgcp62djbr) 20

[Постановка задачі дисертаційного дослідження](#_u9i741d15f3i) 20

[**РОЗДІЛ 2 ПЕРВИННА ОБРОБКА ДАНИХ ТА ПОБУДОВА МОДЕЛІ**](#_dmt7ao9tbz6x) **22**

[2.1 Препроцессінг даних](#_dcf41795lfpq) 22

[2.2.1 Обробка категоріальних даних](#_kz3p4cq4lwvl) 22

[2.2.2 Обробка неперервних даних](#_y0eu6gqtlzd9) 23

[2.2.3 Weight-decay](#_mfx89kewzky1) 25

[2.2.3.1 L2 penalty](#_yoadng9ylz8e) 25

[2.2.3.2 L1 penalty](#_fjkuffz280k4) 27

[2.2.4 Відсів викидів](#_r8dgzmek2pr4) 28

[2.2.4.1 Відсів спостережень](#_hb0lfp6g7xoq) 28

[2.2.4.2. Відсів характеристик](#_mqobfnnm2uyd) 29

[2.2 Побудова моделей](#_n9pfvbun44bo) 29

[2.2.1 Логістична регресія](#_xdde4nw1to05) 29

[2.2.2 Дерева рішень](#_irv2qf7zitei) 30

[2.3 Ансамблі алгоритмів](#_utwbjzeedpui) 31

[2.3.1 Загальна ідея](#_tyn864hrav3) 31

[2.3.2 Bagging](#_37jn4kusey71) 33

[2.3.3 Boosting](#_dco17b4osaha) 35

[2.4 Критерії якості класифікаційних моделей](#_31ds4p832au) 36

[2.4.1 Загальна точність моделі](#_ro1628e8yxq2) 36

[2.4.2 Помилки І-го та ІІ-го роду;](#_lc1ju6qiqj2l) 37

[2.4.3 ROC -крива та індекси AUC та GINI.](#_32528hxxj4qg) 37

[2.4.4 Розбиття датасету](#_dhez8chaahzo) 40

[Висновки за розділом 2](#_w0ii2r1k37e4) 43

[**РОЗДІЛ 3 ПОБУДОВА МОДЕЛЕЙ ЗА ФАКТИЧНИМИ ДАНИМИ**](#_vl7w98iexnh8) **44**

[3.1 Збір даних](#_q2wofaafl0kc) 44

[3. 2 Побудова моделі](#_mgplh5xleei) 47

[3.2.1 Оцінка платоспроможності абонентів “Київстар”](#_8lydxdlijvam) 47

[3.2.2 Оцінка платоспроможності абонентів інших операторів за даними компанії “Київстар”](#_pcrlfzbvici8) 50

[3.3 Аналіз роботи моделей](#_gvx2whsphid6) 51

[3.3.1 Аналіз роботи моделі для прогнозування абонентів “Київстар”](#_mv0y4zm8sx7s) 51

[3.3.2 Аналіз роботи моделі для прогнозування платоспроможності абонентів інших операторів за даними компанії “Київстар”](#_r1bgjx9ibawv) 55

[Висновки за розділом 3](#_o698nl633v2u) 58

[**РОЗДІЛ 4 РОЗРОБКА СТАРТАПУ**](#_1bk9n3qi5gbw) **59**

[4.1 Опис проекту](#_5v5vwbras7g5) 59

[4.2 Команда проекту](#_sh8ie2rgs5k9) 59

[4.3 Бізнес модель проекту](#_kfdkeav9h3iv) 61

[4.4 Аналіз ринкових можливостей стартап проекту](#_4hxatonw4zy6) 63

[4.5 Розроблення ринкової стратегії проекту](#_cu3be5qw26pw) 72

[4.6 Розроблення маркетингової програми стартап-проекту](#_5poi6riqwbg3) 75

[Висновки за розділом 4](#_91vh437nak80) 80

[**ВИСНОВКИ ЗА РОБОТОЮ ТА ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ**](#_soeyepmkwj01) **81**

[**ПЕРЕЛІК ПОСИЛАНЬ**](#_svn4mrnahy34) **83**

[**Додаток А Код програмного продукту**](#_ftrk0qoip1ts) **86**

[**Додаток Б Акт впровадження**](#_5d3ns4stghk) **92**

[**Додаток В Ілюстративні матеріали доповіді**](#_fqjz8u2ycte3) **93**

# 

# ПЕРЕЛІК ПРИЙНЯТИХ ПОЗНАЧЕНЬ ТА СКОРОЧЕНЬ

# 

# 

БС - базова станція

ММП - метод максимальної правдоподібності

МНК - метод найменших квадратів

AUC - Area Under the Curve (площа під кривою)

CA - common accuracy

Bi - відсоток всіх «поганих» випадків у вибірці

FN – false negative

FP - false positive

FPR - False Positives Rate

CEO -Chief Executive Officer

Gi - відсоток всіх «хороших» випадків у вибірці

GINI - коефіцієнт Джини

ROC - receiver operating characteristic

Se - чутливість моделі

Sp - специфічність моделі

TN - true negative

TP - true positive

TPR - True Positives Rate

WOE - Weight of Evidence

# ВСТУП

# 

Для побудови надійного кредитного портфеля кожному банку необхідно постійно вдосконалювати свої скорингові моделі. Для українських фінансових установ основними ресурсами для побудови таких моделей є:

1. Демографічні показники: вік, стать, національність, місце проживання, тривалість проживання в актуальному місці, освіта, професія, тривалість працевлаштування, наявність власності, сімейний стан, наявність дітей і інше. Недоліком такого джерела інформації є відсутність достовірності вказаних клієнтом даних.
2. Фінансова історія клієнта: як часто в минулому клієнт повертав кредит, яка сума надходжень була на його рахунки. Але таке джерело можна використовувати лише у випадку повторного приходу клієнта у банк.

Але як можна покращити такі моделі? Як діяти у випадку відсутності в клієнта кредитної історії. Для розв`язання таких задач необхідно шукати додаткові джерела інформації. Наприклад дані телеком-оператора. Саме цій темі присвячені дослідження моєї магістерської дисертації.

В результаті проведених досліджень за даними телекомунікаційного оператора “Київстар” були побудовані моделі оцінки платоспроможності абонентів як і абонентів “Київстар”, так і абонентів інших операторів.

Результати проведених досліджень були впроваджені в компанії “Київстар”.

# 

# РОЗДІЛ 1 ОЦІНКА ПЛАТОСПРОМОЖНОСТІ КЛІЄНТА

## 1.1 Сутність кредитного ризику і методи боротьби з ним

## 

Кредитний ризик ([англ.](https://uk.wikipedia.org/wiki/%D0%90%D0%BD%D0%B3%D0%BB%D1%96%D0%B9%D1%81%D1%8C%D0%BA%D0%B0_%D0%BC%D0%BE%D0%B2%D0%B0) *credit risk*) – наявний або потенційний ризик для надходжень і [капіталу](https://uk.wikipedia.org/wiki/%D0%9A%D0%B0%D0%BF%D1%96%D1%82%D0%B0%D0%BB), який виникає через неспроможність сторони, що взяла на себе зобов'язання, виконати умови будь-якої фінансової угоди із [банком](https://uk.wikipedia.org/wiki/%D0%91%D0%B0%D0%BD%D0%BA) або в інший спосіб виконати взяті на себе зобов'язання[1].

У банківській сфері при управлінні кредитними ризиками одним з ключових завдань є оцінка кредитоспроможності позичальників. Результати оцінки індивідуальних ризиків є основою для аналізу ризиків усього кредитного портфеля[2].

Оцінка ризику неповернення кредиту по конкретному позичальнику на практиці здійснюється в рамках двох основних підходів - на основі суб'єктивного висновки експертів або на основі автоматизованих систем оцінки кредитного ризику.

Суб'єктивний підхід ґрунтується на суб'єктивних відчуттях, очікуваннях, досвіді осіб, які приймають рішення. Цей підхід має тільки одну перевагу — відсутність витрат на обґрунтування рішення[3]. Ризик при цьому максимальний, а ступінь обґрунтованості рішення, відповідно, мінімальна. Слід зазначити, що у чистому вигляді цей підхід майже не застосовується.

Найбільшого поширення в банківській сфері отримали спеціальні системи з викорисінформація танням автоматизованих моделей для аналізу кредитоспроможності клієнтів. Основною інформацією для таких систем оцінки можуть слугувати про соціально демографічне становище клієнта, кредитної історії, параметрах цього кредиту На даний момент банки висувають підвищені вимоги до ризик-аналітики в зв'язку зі збільшенням випадків шахрайства і зростанням числа безповоротних кредитів[4].

У більшості комерційних банків систем оцінки кредитного ризику є власними розробками з різними методиками на основі даних про позичальників конкретного банку минулих років, або є готовими рішеннями спеціалізованих фірм на основі даних про позичальників декількох банків або фінансових інститутів. І в першому і другому випадку методики систем оцінки кредитного ризику, як правило, становлять комерційну таємницю[5].

Основним етапом побудови моделі є вибір і аналіз незалежних змінних. Прикладі джерел даних:

1. Демографічні показники: вік, стать, національність, місце проживання, тривалість проживання в актуальному місці, освіта, професія, тривалість працевлаштування, наявність власності, сімейний стан, наявність дітей і інше.
2. Дані з приводу причини кредитування: мета кредиту, загальний розмір позики, термін фінансування, початковий внесок, відношення розміру позики до різних міри забезпечення по кредиту та ін.
3. Фінансові показники: при обліку кількісних характеристик в моделі рекомендується уникати абсолютних величин і використовувати коефіцієнти, як наприклад: сума заборгованості до доходу, щомісячні виплати по кредиту до щомісячного доходу, місячний вільно наявний бюджет до щомісячного доходу, щомісячні виплати по кредиту до місячного вільно доступного бюджету та інше.
4. Маркетингові показники: джерело надходження кредитної анкети, проведена програма, що спонукає мотив і інше.

Основним джерелом інформації є внутрішня кредитна історія банку і інформація, отримана в бюро кредитних історій на момент подачі анкети. Використовуваними змінними можуть бути: кількість поточних рахунків клієнта, кількість та наявність кредитних карт, загальна сума всіх кредитів, час отримання останнього кредиту, наявність у клієнта інших продуктів цієї фінансової організації, стан поточного рахунку, утилізація існуючих лімітів, рейтинги бюро кредитних історій та інше[6].

Але такий підхід працює не в усіх випадках. Наприклад компанії, які займаються швидким кредитуванням не можуть зібрати достатню кількість даних про клієнта для того, щоб побудувати гарну модель або це зовсім новий клієнт, який вперше бере кредит в банку й про нього ще нема достатньої кількості інформації. Для таких випадків або для поліпшення роботи вже існуючої моделі (додавання нової характеристики) можна використовувати моделі побудовані іншими організаціями, які мають необхідний вихід на інший канал даних про клієнта.

## 1.2 Нові джерела даних

### 1.2.1 Причини залучення нових даних

## 

Кредитування приватних осіб в Україні знову набирає силу. Товарний кредит на техніку, мікропозики "до зарплати", гроші на освіту і ремонт - до всього цього фінансові установи стають все більш лояльні. За даними НБУ, за перші 4 місяці 2017 року банки видали близько 8 млрд грн нових споживчих кредитів, включаючи карткові позики і іпотеку.

Але в сучасних умовах українська економіка має дуже нестабільний характер. "Наша активність в кеш-кредитуванні та кредитні картки призводить до того, що традиційні методи оцінки фізичних осіб експертним шляхом втрачають ефективність. Тому ми знаходимося в активній фазі розширення застосування скорингових моделей в роздрібному кредитуванні ", - повідомила Ярослава Титова, прес-секретар Ощадбанку. Тому всі фінансові організацію намагаються покращити свої моделі оцінки платоспроможності клієнтів. Для цього вони намагаються знайти нові ресурси даних.

### 1.2.2 Дані соціальних мереж

Даний ресурс є дуже інформативним. З його допомогою можна дізнатися доволі багато інформації про клієнта, а саме:

1. Аналіз контенту клієнта. В якості характеристик для даного аналізу можуть виступати групи на які підписана особа, семантичний аналіз тексту постів, аналіз фотографій та інше. Ці характеристики дають можливість оцінити політичні погляди людини, її соціальний статус, її сферу інтересів та інше.
2. Круг спілкування клієнта. Яку кредитну історію має його круг спілкування [7].
3. Кількість друзів в соціальній мережі
4. Стиль життя людини. З допомогою технологій Computer Vision можна провести аналіз фотографій клієнта та виявити чи веде він не здоровий спосіб життя, чи займається екстремальними видами спорту. Усі ці показники зменшують ймовірність видачі кредиту, оскільки у випадку[8]

Але соціальні мережі мають один великий недолік. а саме наявність фейкових сторінок. Для цього необхідно дивитися на кількість друзів, на кількість лайків, на кількість фейкових друзів (оскільки в соціальних мережах часто використовують спеціальні програми для збільшення друзів та лайків), на тривалість існування сторінка, оскільки чим довше існую сторінка тим більша ймовірність, що дана сторінка не фейк та інше [9].

В таблиці 1.1 продемонстрований приклад хорошого та поганого клієнта:

Таблиця 1.1–Приклад хорошого та поганого клієнта за даними соціальної мережі

|  |  |  |
| --- | --- | --- |
| Критерій | Хороший клієнт | Поганий клієнт |
| Дата створення сторінки | 2012 | 2017 |
| Аналіз зображень, які викладає клієнт | Сімейні цінності | Аморальний контекст |
| Кількість друзів | 300 | 20 |
| Друзі | Мають великий середній скоринг | Мають малий середній скоринг |
| Адреса електронної пошти | Smirnov@gmail.com | TJRNFHNERHJ@mail.ru |
| Семантичний аналіз текстів | Не використовує нецензурну лексику | Використовує нецензурну лексику |

Соціальні мережі є постачальниками інформації і для кредитних бюро в США. Їх співробітники вивчають фотографії потенційного позичальника, наприклад, як часто людина фотографується в п'яному вигляді. Шанси нетверезих веселунів на отримання кредиту зменшуються. Крім того, аналітики звертають увагу на те, в які магазини і як часто ходить людина, які ЗМІ читає, який соціальний статус його друзів, в якому навчальному закладі він навчався і хто його однокурсники та однокласники, на яку інформацію ставить like. Ці дані дозволяють скласти психологічний портрет позичальника і важливі не менше, ніж фінансова інформація про розмір зарплати і кількості боргів.

Компанія Kreditech з Німеччини надає мікропозики терміном до 30 днів, грунтуючись на власній скорингової моделі, яка включає в себе і аналіз активності позичальника в соціальних мережах. Всього алгоритм скорингу враховує 8000 різних чинників, зазначає глава компанії Себастьян Даймер. Крім соціальних мереж Kreditech аналізує покупки і інші дії позичальника в інтернеті, за якими можна судити про його перевагах і способі життя. При можливості визначається і його місцезнаходження. Все це дозволяє відсіяти потенційних шахраїв. У компанії заявляють, що прострочення по кредитах становить менше 10%.

Серед першопрохідців у країнах СНГ є Альфа-банк, який впровадив таку модель у 2015 році, “Тінькоф”, який для своїх моделей використовує інформацію про те наскільки давно зареєстрований користувач, кількість його друзів та активність

### 1.2.3 Дані “Яндекс Таксі” або Uber

За допомогою таких даних можна дізнатися інформацію як і про людей, які викликають таксі, так і про самих водіїв.

Наприклад про пасажирів можна сказати:

- як часто клієнт використає послуги таксі;

- які відгуки про клієнта залишають водії ;

А про водіїв:

- який середній бал від пасажирів вони отримали;

- скільки поїздок в них було;

- чи не потрапляли вони ДТП;

Компанія Uber одною з перших, хто почав запускати подібний продукт в країнах зі слабкою кредитною інфраструктурою. Серед країн першопрохідців була Кенія. Поки місцеві кредитні організації не могли визначити кредитоспроможність клієнтів через нестачу інформації, кенійський Sidian Bank почав використовувати інформацію Uber про таксистів. Існуючі дані стали підставою для видачі кредитів на покупку особистого авто. Сервіс почав працювати в травні 2016.

“Замість традиційних банківських методів, ми довіряємо даним отриманим від Uber” - заявив генеральний директор Sidian Bank Титус Каранжа

Щоб отримати кредит від банку, таксист Uber повинен відповідати певним критеріям: накопичити понад 500 поїздок і середній рейтинг пасажирів не менше 4,6 з 5.

Треба відмітити, що через відсутність даних і кредитної історії в Кенії з населенням 45 млн чоловік тільки 4,4% жителів отримували позики. Банки не хочуть ризикувати.

### 1.2.4 Дані про поведінку користувача на сайті банку

Також доволі цікавим ресурсом даних є поведінка користувача на сайті компанії. Але таке джерело доступно лише у випадку онлайн кредитування.

Основними характеристиками, які використовуються для побудови моделей є:

1. Пристрій з якого заходить клієнт. i-Phone, Samsung чи просто кнопковий телефон.
2. Чи була затримка у прийняті рішень під час заповнення анкети ? Якщо так, то при заповнення яких саме. Наприклад, якщо користувач довго думав, який саме рівень свого доходу вказати й спочатку заповняв дане поле одним значенням, а потім почав його змінювати, то така поведінка може вказати про те, що даний клієнт давав спеціально невірну інформацію.

Серед компаній, які вже використовують такі технології входить: MoneyMan, AmmpPay та інші.

### 

### 

### 1.2.5 Дані телеком-оператора

Саме цьому джерелу даних й присвячена дана робота. Цей ресурс дозволяє дізнатися про:

- геолокацію клієнта[10];

- його інтернет трафік;

- його активність у мережі.

На українському ринку одним з першопрохідців у даній сфері була компанія “Київстар”. Як повідомив CEO компанії П.Чернишов на сторінці в Фейсбуці, Альфа-Банк і роздрібний мобільний банк Monobank вже використовують скоринг від "Київстар".

В таблиці 1.2 продемонстрований приклад хорошого та поганого клієнта:

Таблиця 1.2 – Приклад хорошого та поганого клієнта за даними телеком-оператора

|  |  |  |
| --- | --- | --- |
| Критерій | Хороший клієнт | Поганий клієнт |
| Дата активації мобільної картки | 2012 | 2017 |
| Кількість дзвінків без відповіді | 3 | 20 |
| Кількість поїздок за кордон протягом останніх 6 місяців [11] | 2 | 0 |
| Тарифний план | Найдорожчий | Часто має 0 на рахунку |
| Модель телефону[12] | iPhone 7 | Кнопковий телефон |
| Змінює свою локацію протягом робочих днів | Так | Ні |

Окрім звичайного скорингу, якщо клієнт дав згоду на обробку персональних даних телекомунікаційні оператори можуть запропонувати таку послугу, як “тригер”. Її ідея полягає в тому, що коли позичальник змінює місце свого проживання або перестає користуватися послугами телеком-оператора у банкам відправляється повідомлення про можливу загрозу пов’язану з ухиленням кредитора від своїх зобов’язань[13].

## 

## Висновки до розділу 1

В даному розділі були розглянуті основні принципи боротьби з кредитним ризиком. Описані стандартні джерела даних, які використовуються для побудови якісного кредитного портфелю, а саме: кредитна історія клієнта, анкетні дані, історія фінансових надходжень на корту.

Також сформована проблема неповноти даних про клієнта, та описані сучасні методи подолання даної проблеми, які представляють собою залучення нових ресурсів даних, таких як: дані з соціальних мереж, дані телекомунікаційного оператора, дані про поведінку клієнта на сайті під час заповнення онлайн анкетування та інші.

Таким чином, ставиться задача підвищення якості оцінювання кредитоспроможності клієнтів фінансових установ. Для цього необхідно використовувати додаткові джерела інформації, будувати спеціалізовані системи обробки даних та генерування скорингу, що і буде розглянуто у наступних розділах дисертаційного дослідження.

## Постановка задачі дисертаційного дослідження

1. Виконати дослідження проблеми підвищення якості скорингового аналізу даних клієнтів-позичальників кредитів фінансових установ.
2. Зібрати необхідні статистичні дані і вибрати методи моделювання для оцінювання кредитоспроможності клієнтів.
3. Побудувати моделі для прогнозування платоспроможності абонентів “Київстар”
4. Побудувати моделі для прогнозування платоспроможності абонентів інших операторів за даними компанії “Київстар”
5. Виконати порівняльний аналіз отриманих результатів та запропонувати кращі моделі для практичного використання.

# 

# РОЗДІЛ 2 ПЕРВИННА ОБРОБКА ДАНИХ ТА ПОБУДОВА МОДЕЛІ

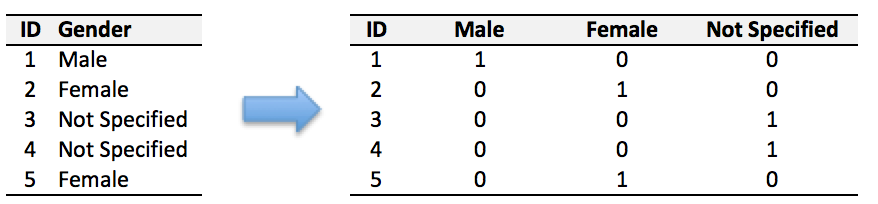
## 2.1 Препроцессінг даних

### 2.2.1 Обробка категоріальних даних

One Hot Encoding

Основна ідея методу

Припустимо, що деяка ознака може приймати 10 різних значень. В цьому випадку One Hot Encoding має на увазі створення 10 ознак, всі з яких дорівнюють нулю *за винятком одного* . На позицію, відповідну чисельним значенням ознаки ми поміщаємо 1. Даний алгоритм наведений на рисунку 2.1.

Рисунок 2.1 – One Hot Encoding 

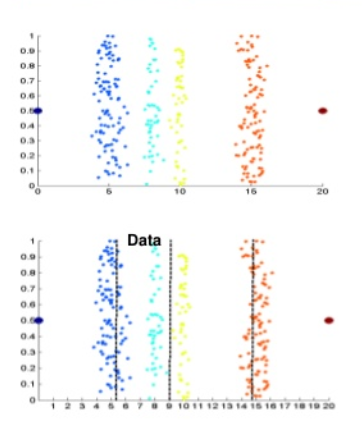
Цей метод реалізовано в бібліотеці sklearn.preprocessing в класі OneHotEncoder. За замовчуванням OneHotEncoder перетворює дані в розріджену матрицю, щоб не витрачати пам'ять на зберігання численних нулів.

### 2.2.2 Обробка неперервних даних

Бінінг

Основна ідея методу

По кожній з неперервних характеристик ми будуємо своє дерево рішень для прогнозу цільового значення. Кожна гілка представляє собою нову категоріальну характеристику. Бажано будувати не глибокі дерева (до 5-6 гілок) для відсутності перенавчання[14]. Реалізація бінінгу наведена на рисунку 2.2.

Рисунок 2.2 – Реалізація бінінгу

Коли використовується?

Застосовується для задач класифікації. Оскільки в результаті використання даного методу прогнозоване значення буде мати дискретне значення (бо усі вхідні характеристики також дискретні).

Плюси при застосуванні

1. Модель при застосуванні бінінгу стає менш чутлива до змін. З часом в деяких характеристиках може змінюватись їх математичне очікування. Наприклад показники середніх витрат на рахунку з часом може збільшуватись за рахунок інфляції. За рахунок перетворення неперервної характеристики на категоріальну ми більше застраховані до подібних змін.
2. Обробка пропусків. Це особливо відчутно при великому обсязі характеристик в нас немає необхідності обробляти придумувати особливий спосіб заповнення пропущених значень, оскільки ми можемо завести для таких значень окремий бін.
3. Боротьба з викидами. Якщо в якійсь харктерстиці з`явиться не притаманне для неї велике або мале значення, то модель не буде перенавчатись на цьому окремому прикладі, оскільки наша характеристика потрапить до окремого біна, який буде символізувати велике або мале значення
4. Чутливість до нелінійних залежностей[15].

Припустимо, що в нас є показник «вік». Значення таргету має сильний зв`язок з особами вік яких складає від 22 до 30. Якщо в якості основної моделі ми використовуємо наприклад логістичну регресію, то дана характеристика має свою певну вагу, що значить, що при збільшені цього показника ми створюємо більший вплив на таргетне значення, а при зменшенні навпаки зменшуємо, але таргетне значення найбільше залежить для показників, які знаходяться посередині. Якщо ж ми застосовуємо бінінг, то в нас з`явится нова характеристика, яка буде символізувати, що вік особи складає від 22 до 30 й при наявності такої характерстики в нас буде посилюватися вплив на таргетне значення.

Недоліки при застосуванні

1. Втрата інформації.

### 2.2.3 Weight-decay

#### 2.2.3.1 L2 penalty

Стандартний штраф за вагу L2 включає додавання додаткової змінної до функції вартості. Це робить ваги невеликими, якщо вони не мають великих похідних помилок[16].

, (2.1)

де , (2.2)

, (2.3)

, \* , (2.4)

де - параметр регуляризації;

;

y - результат роботи моделі;

t - значення, яке прогнозується;

С - цільова функція.

Це запобігає використанню моделлю ваг, які йому не потрібні. Це робить модель менш чутливою, в якій вихід змінюється повільніше при зміні входу.

Якщо в моделі є два дуже схожих входу даних, вона вважає за краще покласти половину ваги на кожен, а не всю вагу на один. Це зображено на рисунку 2.3.

### 

Рисунок 2.3 – Основна ідея l2 регуляризації

#### 2.2.3.2 L1 penalty

Іноді краще працює l1 penalty, щоб оштрафувати абсолютні значення ваг. Це може зробити багато ваг в точності рівними нулю, що дуже допомагає тлумаченню роботи моделі та відсіву слабких характеристик[17].

Якщо подивитися на результати експерименту наведеного нижче, то чітко видно, що при додавання несуттєвих характеристик до моделі при використанні l1 penalty її помилка майже не змінюватиметься, на відміну від використання l2 penalty [4]. Приклад такої ситуації зображений на рисунку 2.4.

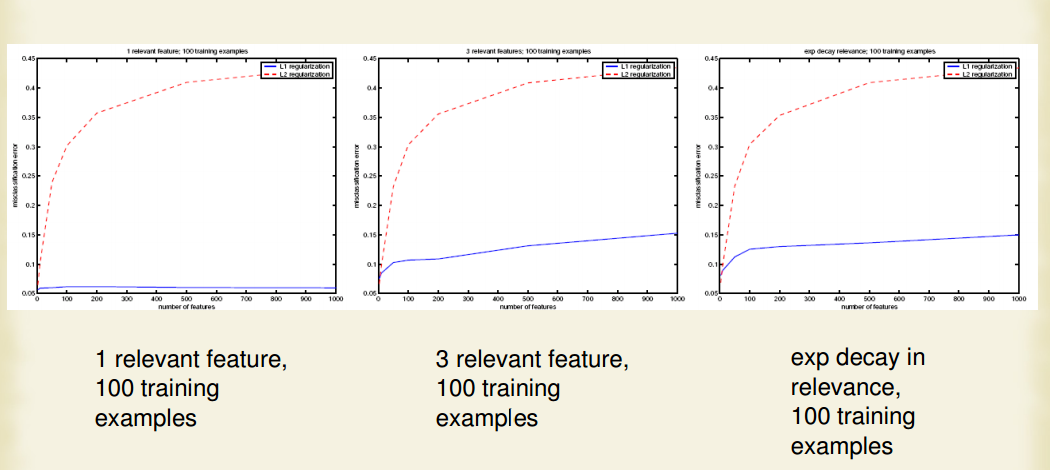


Рисунок 2.4. Робота l2 регуляризації

### 

### 2.2.4 Відсів викидів

#### 2.2.4.1 Відсів спостережень

OneClassSVM

Основна ідея:

Побудова опорної векторної машини для виявлення 1-го класа. Все, що не входить до цього класу є викидами [14]. Приклад роботи OneClassSVM зображений на рисунку 2.5.

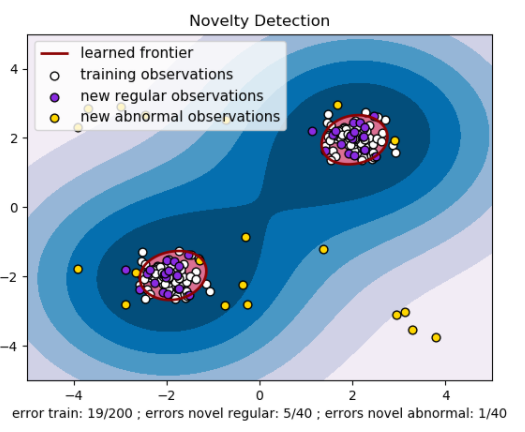


Рисунок 2.5 – One Class SVM

#### 2.2.4.2. Відсів характеристик

Відсів сильно коррелюючих змінних

Основна ідея

Якщо в модель входить 2 сильно коорелюючі характеристики (модуль корреляціії більший за 0.7), то бажано відкинути одну з них, оскільки це може слугувати приводом для перенавчання моделі.

## 2.2 Побудова моделей

### 2.2.1 Логістична регресія

Основна ідея:  
 Логістична регресія - найпоширеніша статистична модель для побудови скорингових карт при бінарно залежній змінній[15]. Математично модель логістичної регресії виражає залежність логарифма шансу від лінійної комбінації незалежних змінних:

, (2.5)

де - ймовірність настання дефолту по кредиту для i-го позичальника; значення j-ой незалежної змінної; - незалежна константа моделі, – параметри моделі;- компонент випадкової помилки.

У логарифмічній регресії скорингові бали трансформовані в ймовірності зі значеннями від 0 до 1.

### 2.2.2 Дерева рішень

Основна ідея

Дерева рішень - це модель, яка будується на логічному ланцюжку правил, які намагаються описати окремі взаємозв'язку між даними відносно очікуваного результату[16]. Структура дерев рішень відкрито відображає аргументацію правил і тому дозволяє легко зрозуміти процес прийняття рішення[17]. Графічно спрощена модель показана на рисунку 2.6.

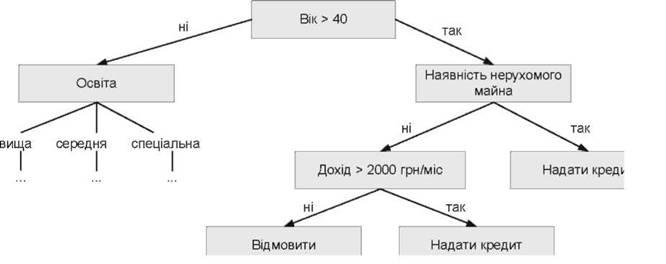


Рисунок 2.6 – Дерево рішень

## 2.3 Ансамблі алгоритмів

### 2.3.1 Загальна ідея

Хорошим прикладом ансамблів вважається теорема Кондорсе «про журі присяжних» (1784). Якщо кожен член журі присяжних має незалежну думку, і якщо ймовірність правильного рішення члена журі більше 0.5, то тоді ймовірність правильного рішення присяжних в цілому зростає зі збільшенням кількості членів журі і прагне до одиниці. Якщо ж ймовірність бути правим у кожного з членів журі менше 0.5, то ймовірність прийняття правильного рішення присяжними в цілому монотонно зменшується і прямує до нуля зі збільшенням кількості присяжних.

(2.6)

де N — кількість присяжних  
p — ймовірність правильного рішення присяжного  
μ — ймовірність правильного рішення всього журі

Repeat{

:= -

}

Якщо p > 0.5, то μ > p. Якщо , то

Ще один популярний приклад ансамблів - "Мудрість натовпу". Френсіс Гальтон в 1906 році відвідав ринок, де проводилася якась лотерея для селян.

Їх зібралося близько 800 чоловік, і вони намагалися вгадати вагу бика, який стояв перед ними. Бик важив 1198 фунтів. Жоден селянин не вгадав точну вагу бика, але якщо порахувати середнє від їх прогнозів, то отримаємо 1 197 фунтів.

Цю ідею зменшення помилки застосували і в машинному навчанні.

Основні способи отримання середнього по моделям:

1. Арифметичне середнє. Ми можемо об'єднати декілька моделей взявши арифметичне середнє їх прогнозів [18]. Приклад арифметичного середнього зображений на рисунку 2.7.

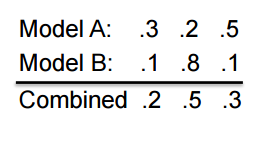


Рисунок. 2.7–Арифметичне середнє

1. Геометричне середнє. Ми можемо об'єднати декілька моделей взявши геометричне середнє їх прогнозів. Приклад роботи геометричного середнього зображений на рисунку 2.8.

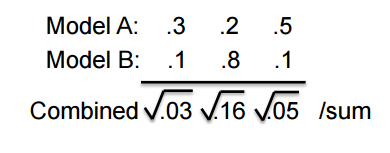


Рисунок.2.8 – Геометричне середнє

### 2.3.2 Bagging

Bagging (від Bootstrap aggregation) - це один з перших і найпростіших видів ансамблів. Він був придуманий Лео Брейманом в 1994 році. Беггінг заснований на статистичному методі бутстрепа, який дозволяє оцінювати багато статистики складних розподілів.

Основною ідеєю алгоритма є паралельне навчання класифікаторів.

Метод бутстрепа полягає в наступному. Нехай є вибірка X розміру N. Рівномірно візьмемо з вибірки N об'єктів з поверненням. Це означає, що ми будемо N раз вибирати довільний об'єкт вибірки (вважаємо, що кожен об'єкт «дістається» з однаковою ймовірністю 1 / N), причому кожен раз ми вибираємо з усіх вихідних N об'єктів. Можна уявити собі мішок, з якого дістають кульки: обраний на якомусь етапі кулька повертається назад в мішок, і наступний вибір знову робиться рівноможливо з того ж числа кульок. Відзначимо, що через повернення серед них виявляться повтори. Позначимо нову вибірку через X1. Повторюючи процедуру M раз, згенеруємо M підвибірок X1, ..., XM. Тепер ми маємо досить велику кількість вибірок і можемо оцінювати різні статистики вихідного розподілу[19]. Приклад роботи бегингу наведений на рисунку 2.9.

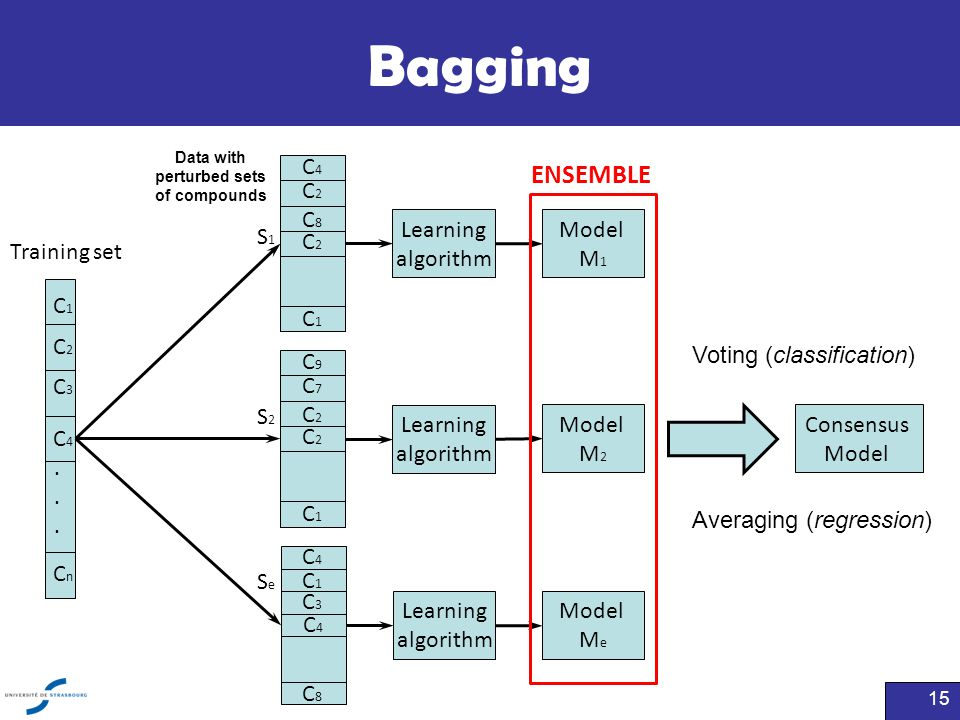


Рисунок.2.9 – Bagging

Найбільш відомим bagging алгоритмом є Random forest.

Алгоритм машинного навчання, запропонований Лео Брейманом і Адель Катлер, що полягає у використанні комітету (ансамблю) вирішальних дерев. Алгоритм поєднує в собі дві основні ідеї: метод беггінга Брейман і метод випадкових підпросторів, запропонований Tin Kam Ho. Алгоритм застосовується для задач класифікації, регресії і кластеризації.

Основна ідея

RF (random forest) - це безліч дерев рішень. У задачі регресії їх відповіді усереднюються, в завданню класифікації приймається рішення голосуванням за більшістю. Всі дерева будуються незалежно за схемою, поданою нижче.

Вибирається підвибірка навчальної вибірки розміру samplesize (м.б. з поверненням) - по ній будується дерево (для кожного дерева - своя підвибірка).

Для побудови кожного розщеплення в дереві переглядаємо max\_features випадкових ознак (для кожного нового розщеплення - свої випадкові ознаки).

Вибираємо найкращі ознаки і розщеплення по ньому (по заздалегідь заданому критерію). Дерево будується, як правило, до вичерпання вибірки (поки в листі не залишаться представники тільки одного класу), але в сучасних реалізаціях є параметри, які обмежують висоту дерева, число об'єктів в листі і число об'єктів в підвибірки, при якому проводиться розщеплення [20].

### 2.3.3 Boosting

Cуть алгоритму

В основі алгоритму лежить почергове навчання моделей. Наприклад перший бустінговий класифікатор Boost1 складався з 3-х класифікаторів, перший класифікатор навчався на всьому наборі даних. другий на іншому наборі, а третій навчався на тих прикладах де 1-й та 2-й класифікатори розійшлися у відповдях [21].

,

, (2.8.)

, (2.9.)

). (2.10)

Раннє використання boosting проходило з нейронними мережами для MNIST.

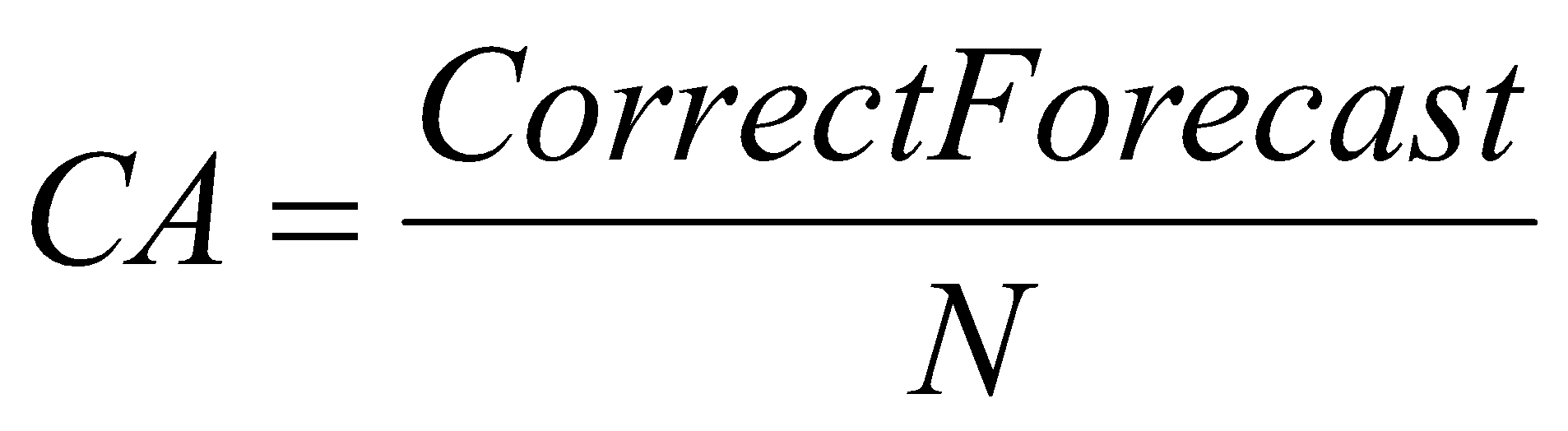
Найбільш популярні boosting алгоритми:

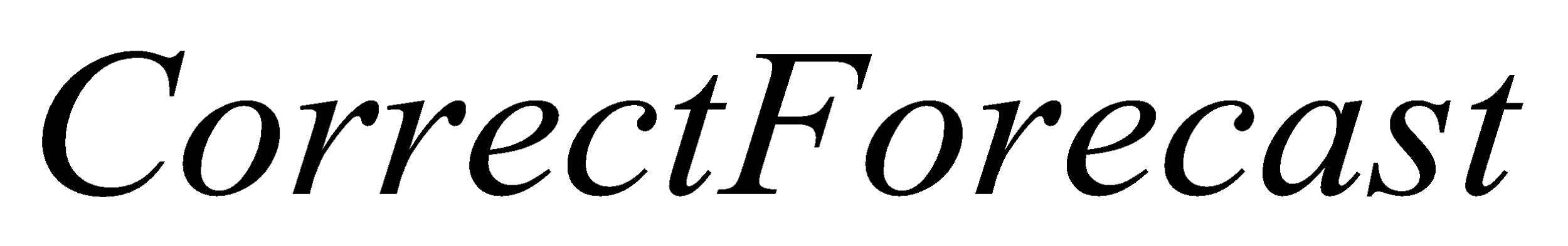
* AdaBoost (адаптивне підвищення) ;
* Gradient boosting.

## 2.4 Критерії якості класифікаційних моделей

### 2.4.1 Загальна точність моделі

Загальна точність моделі (CA – common accuracy) визначається як:

, (2.11)

де  – кількість вірно спрогнозованих випадків, а – загальна кількість випадків. Загальна точність моделі є дещо суб’єктивною оцінкою, оскільки вона залежить від долі дефолтів у моделі, а також від порогу відсікання. Для різних значень порогу точність моделі також буде приймати різні значення.

### 

### 2.4.2 Помилки І-го та ІІ-го роду;

Основні види помилок наведені в таблиця 2.1.

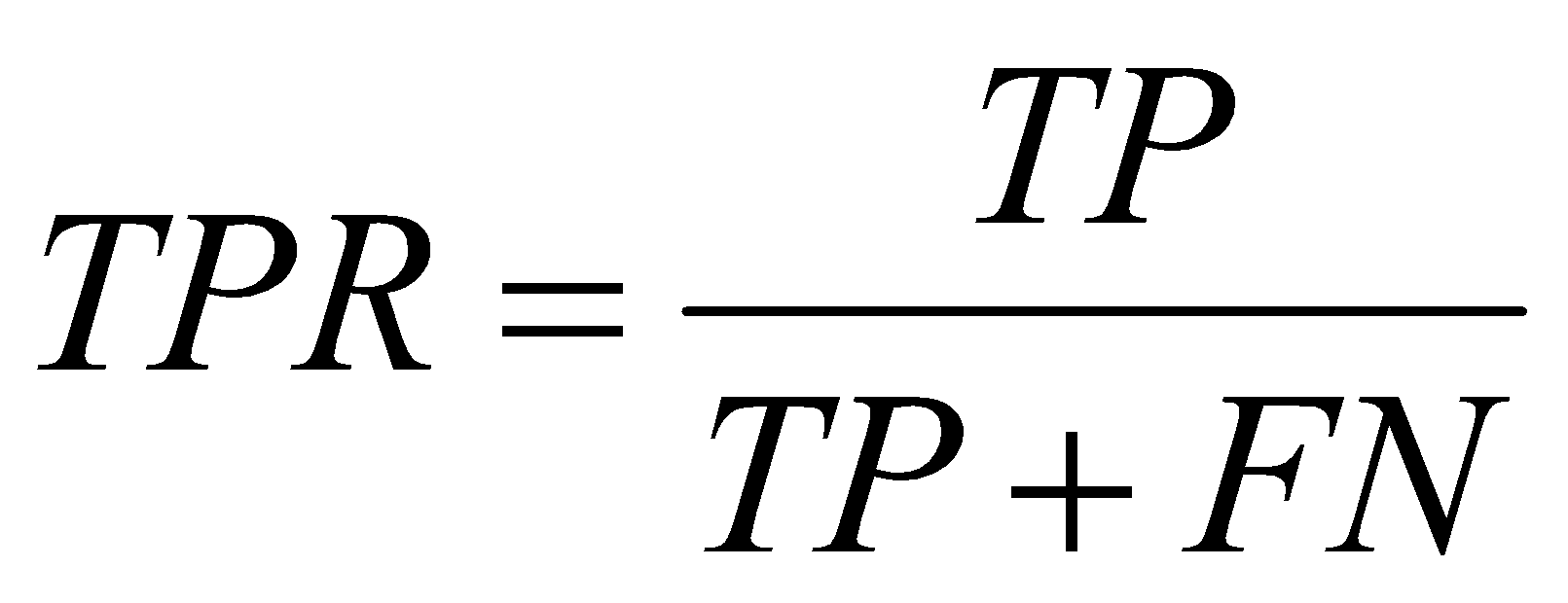
Таблиця 2.1–Помилки І-го та ІІ-го роду

|  |  |  |
| --- | --- | --- |
|  | Прогноз моделі: Повернення кредиту (0) | Прогноз моделі: Дефолт (1) |
| Фактично: Повернення кредиту (0) | Вірно класифіковані (TP) | Помилки ІІ-го роду (FN) |
| Фактично: Дефолт (1) | Помилки І-го роду (FP) | Вірно класифіковані (TN) |

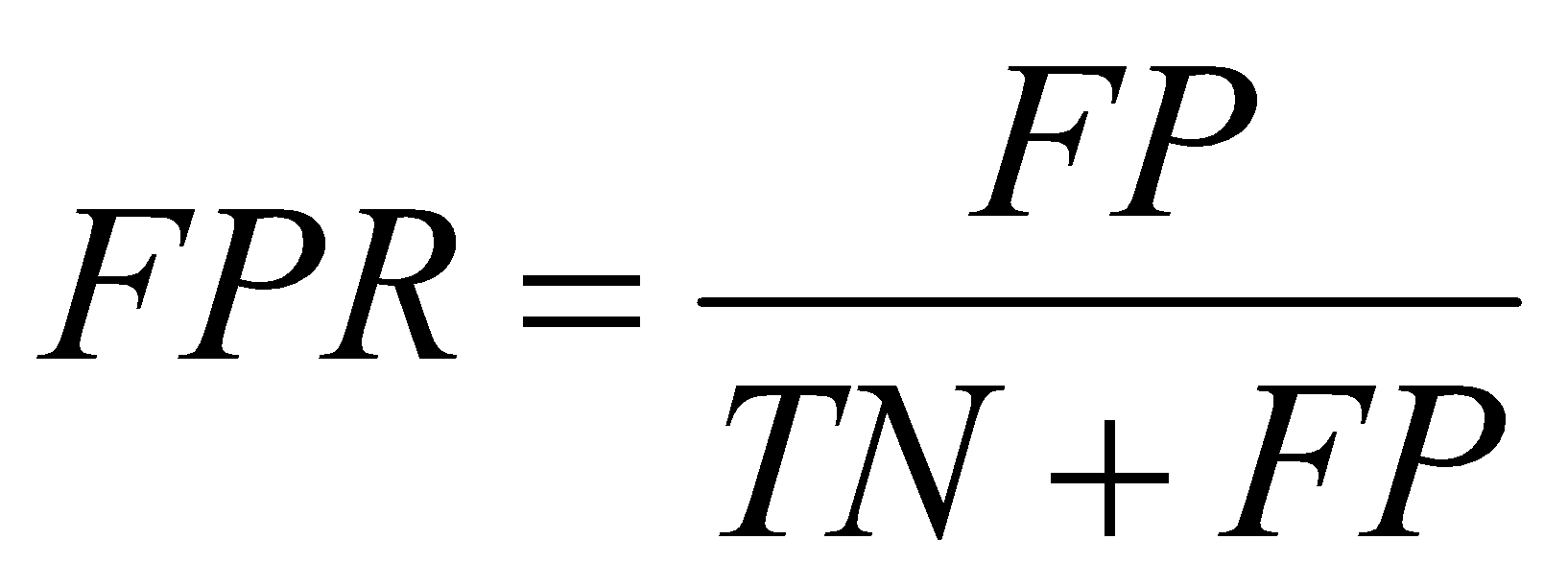
### 2.4.3 ROC -крива та індекси AUC та GINI.

Відносні показники у процентах:

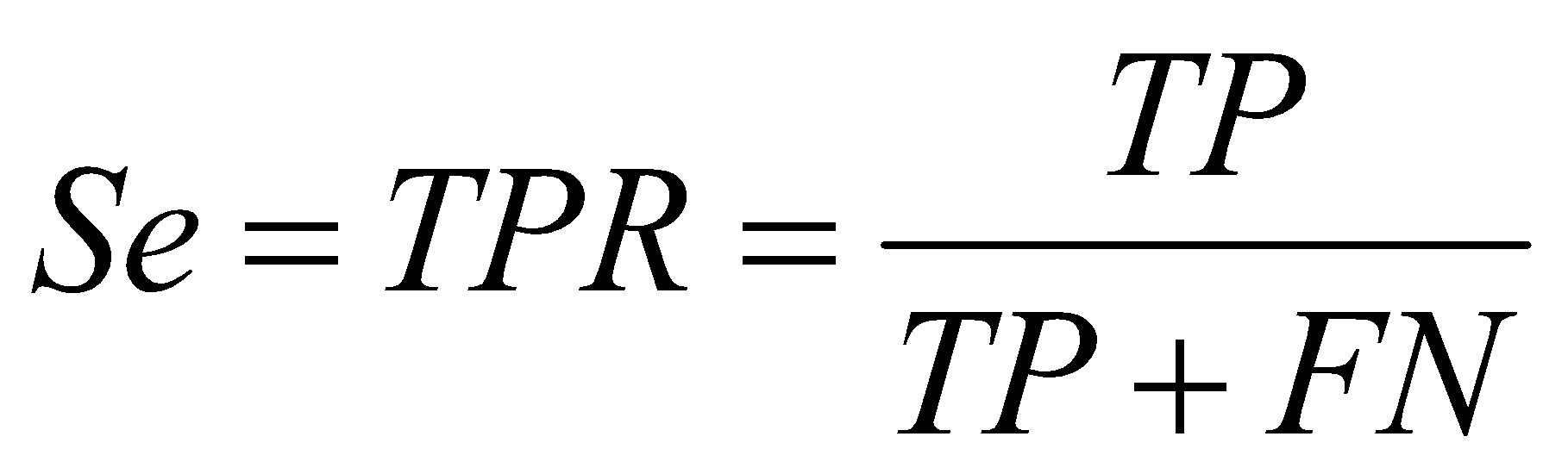
а) частка істинно позитивних прикладів (True Positives Rate):

 (2.12)

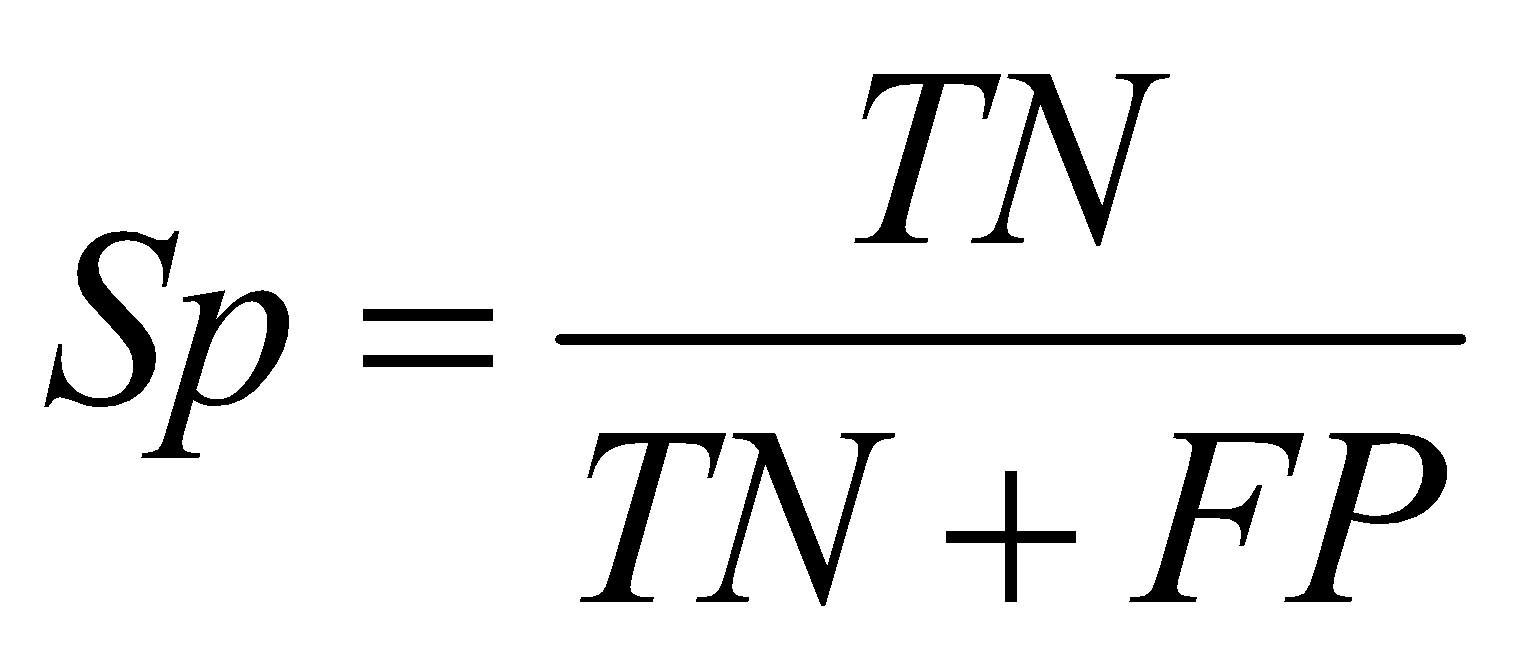
б) частка хибно позитивних прикладів (False Positives Rate):

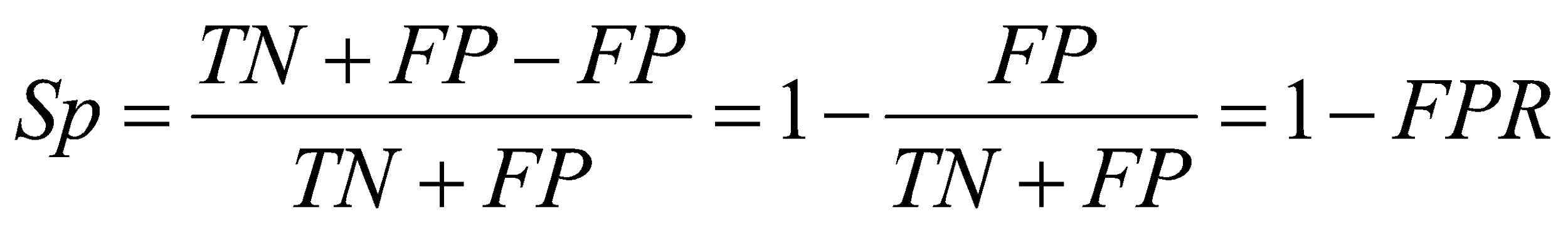
 (2.13)

в) чутливість моделі – це частка істинно позитивних випадків, тобто:

 (2.14)

г) специфічність моделі – це частка істинно негативних випадків, які були вірно класифіковані моделлю:

 (2.15)

 (2.16)

Для оцінки якості класифікації часто вдаються до аналізу ROC-кривих (від англ. Receiver Operator Characteristic). Цей аналіз також дозволяє вибрати оптимальний поріг ймовірності поділу «поганих» і «хороших» позичальників для досягнення прийнятного рівня чутливості і специфічності моделі [22]. Приклад ROC кривої зображено на рисунку 2.10.

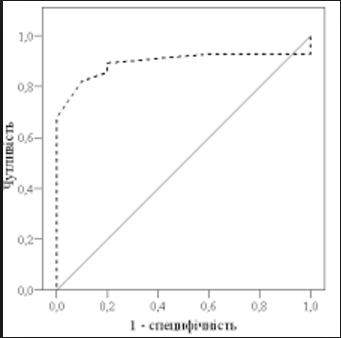
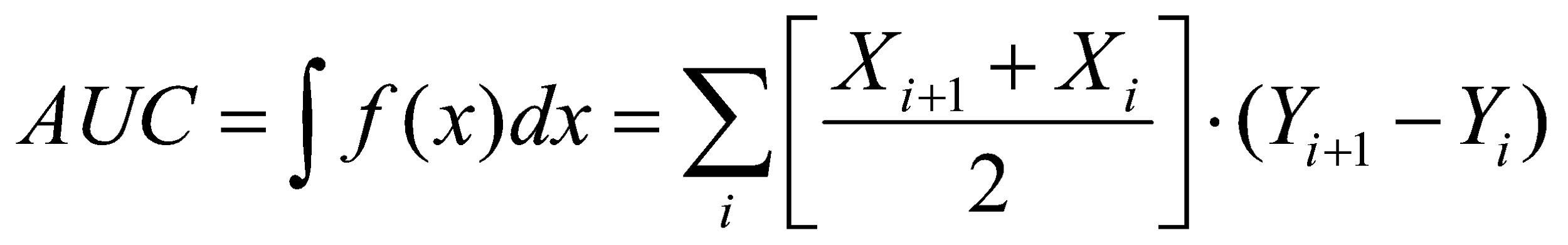


Рисунок 2.10–Графік ROC-кривої

Для аналізу класифікаційної здатності моделі оцінки кредитного ризику використовують показник площі під ROC-кривої - AUC (від англ. Area Under Curve). Площа під кривої AUC змінюється від 0.5 (немає поділу) до 1 (ідеальне поділ). Зазвичай вважають, що значення площі від 0.9 до 1 відповідає відмінній якості моделі, від 0.8-0.9 - дуже хорошому, 0.7-0.8 - хорошому, 0.6-0.7 - середнього, 0.5-0.6 - незадовільного. Слід зазначити, що показник площі під кривою призначений тільки для порівняльного аналізу моделей між собою. Значення площі під кривою не містить ніякої інформації про чутливість і специфічність моделі.

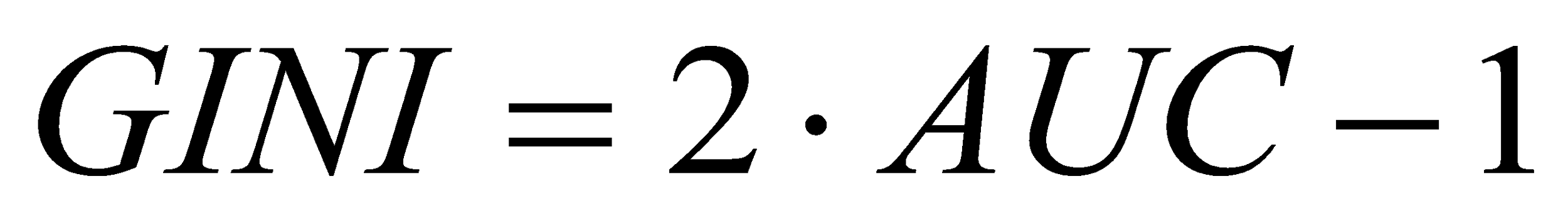
На практиці цей показник часто розраховується за формулою:

 (2.17)

При аналізі якості моделі за значенням площі під ROC-кривої часто обчислюють індекс Джині. Цей показник переводить значення площі під кривою в діапазон від 0 до 1, чим вище його величина, тим вище дискримінуюча здатність моделі.

Індекс GINI – це площа області між діагоналлю і кривою Лоренца, поділена на площу усієї області під діагоналлю. Індекс GINI проілюстровано графом, де сукупний процент дефолту для клієнтів показано поряд із сукупним процентом клієнтів, коли вони упорядковані за ймовірністю дефолту (менша ЙД – зліва, більша - справа). Цей графік відомий як крива Лоренца (CAP-крива). При цьому припускається, що у класифікатора існує певний параметр, варіюючи який, отримуємо певне розбиття на класи. Цей параметр часто називають порогом або точкою відсікання (сut-off), в залежності від якого будуть отримані різні величини помилок І-го та ІІ-го роду.

Розраховується індекс Джині за формулою:

 (2.18)

### 2.4.4 Розбиття датасету

Неправильний метод полягає в тому, щоб спробувати безліч альтернатив і подивитися, що дає кращу продуктивність в тестовому наборі. Це легко зробити, але це дає неправильне уявлення про те, наскільки добре працює цей метод. Налаштування, які найкраще працюють на тестовому наборі, навряд чи будуть працювати також на новому тестовому наборі, взятому з одного і того ж дистрибутива.

*Крайній приклад.* Припустимо, що модель дає безліч варіантів відповіді, які не залежать від характеристик, які подавалися на вхід. Припустимо, що на окремому наборі даних дана модель випадковим чином дає непогані результати. Але не можна очікувати, таких самих показників на новому тестовому наборі.

Які є методи для боротьби с даною проблемою ?

Cross-validaion. Ідея алгоритму полягає у розбиті всього датасету на n частин та подальшому проведені n ітерацій. На кожній ітерації в якості тестового набору вибирається нова i-та частина датасету, а навчання ведеться на інших n-1 частинах. Всі необхідні для аналізу метрики рахуються n разів після чого підраховується їх середнє, мінімальне, максимальне та дисперсія на основі яких вже й будуються певні висновки про якість роботи побудованої моделі [23]. Реалізація показана на рисунку 2.11.

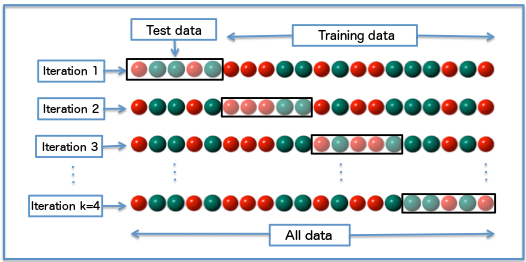


Рисунок 2.11 – Cross-validation

Для зменшення розбіжності у показниках можна також використовувати стратифікацію. Основна ідея полягає в тому, що загальна вибірка розбивається на n підвибірок у яких зберігається процентне співвідношення таргету. Оскільки при такому підході зменшується вірогідність того, що до навчальної вибірки випадковим чином потрапить більше одного чи іншого таргету, що може спричинити перенавчання та в свою чергу не відповідність експерименту до реальної ситуації. Реалізація наведена на рисунку 2.12.

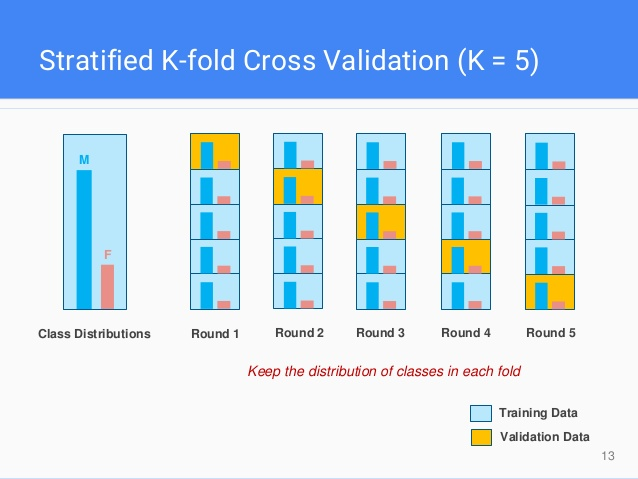


Рисунок 2.12 – Stratified k-fold cross-validation

## Висновки за розділом 2

В даному розділі описані основні методи обробки даних. А саме, основні ідеї обробки неперервних даних, обробки категоріальних даних та відсіву викидів, що необхідно для того щоб побудувати адекватну модель. Описана ідея використання біннінгу для обробки даних при побудові скорингових моделей. Розглянуто відмінність між l1 та l2 регуляризаціэю.

Також розглянуті найбільш популярні моделі для оцінювання кредитоспроможності клієнтів фінансових установ. Це моделі інтелектуального аналізу даних, такі як логістична регресія, дерева рішень, Random forest, градієнтний бустінг. Описана основна ідея бегінгу та бустінгу.

Розглянуті основні критерії, за допомогою яких проводиться оцінка скорингових моделей і встановлюється якість класифікації. А саме, розглянуті такі якісні показники як GINI та AUC, загальна точність моделі, помилка першого та другого роду. Описані основні підходи, які використовуються для коректного тестування моделей, а саме, ідея cross-validation та stratified k-fold cross-validation.

# РОЗДІЛ 3 ПОБУДОВА МОДЕЛЕЙ ЗА ФАКТИЧНИМИ ДАНИМИ

## 3.1 Збір даних

Для побудови моделі взяті реальні данні за бази даних “Київстар”, який є найбільшим українським оператор мобільного зв'язку, який обслуговував станом на кінець третього кварталу 2017 року біля 26,4 млн абонентів.

На початку побудова необхідно було відібрати характеристики, які могли б вплинути на оцінку платоспроможності клієнтів.

Для оцінки номерів телеком оператора були відібрані наступні ознаки:

а) Оплата телеком-послуг. В якості основних характеристик обрано середні затрати на поповнення рахунку, динаміка поповнення телеком послуг (відношення суми оплат за останні місяці до суми оплат за минулі місяці)

б) Інформація про інтернаціональну активність абонента:

1. Затрати на інтернаціональні дзвінки, які можуть свідчити про дохід клієнта.
2. Динаміка подорожей клієнта. Оскільки кожна активність клієнта фіксується на окремій базовій станції можна слідкувати за кількістю відвіданих абонентом країн.

в) Використання кредитних продуктів телекомунікаційного оператора. З рисунку 3.1 видно значимість цієї характеристики.

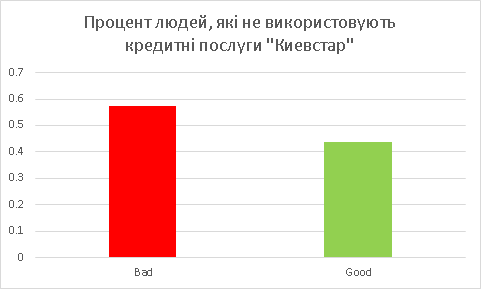


Рисунок 3.1 – Процент людей, які не використовують кредитні послуги “Київстар”

г) У більшості телекомунікаційних операторів є можливість надання абоненту додаткових коштів на дзвінки у кредит:

1. Динаміка взяття кредиту
2. Динаміка закриття кредиту

ґ) Час з моменту підключення. Оскільки часто шахраї змінюють мобільні телефони це може бути доволі інформативною характеристикою.

д) Кількість дзвінків без відповіді. Як не дивно ця характеристика часто може проявити свою значущість. Можливо це пов`язано з тим, що люди, які не повертають кредити часто не відповідають на дзвінки колекторів.

е) Використання інтернет трафіку

є) Популяція населеного регіону. Ця характеристика може сильно корелювати з середнім заробітком.

ж) Чи відрізняються у абонента робоча та домашня зона. Ця характериститка є певною апроксимацією працевлаштованості абонента. Більшість людей з 00:00 до 6:00 знаходяться вдома, а з 8:00 до 18:00 на роботі. За цей час до людини надходять дзвінки, смс або вона використовує інтернет. Всі ці події фіксуються на певній базовій станції через яку надходить сигнал. Якщо знайти базову станцію через яку протягом цього часу проходила найбільша кількість подій, то можемо зробити гіпотезу, що на робочу або домашню зона абонента світить саме ця БС.

з) Чи працює абонент у вихідні дні. Таке твердження можна перевірити тим самим методом, як і в минулому пункті, а саме. якщо по минулій характеристиці стоїть 1, то перевірити чи знаходиться абонент у вихідні дні у зоні на яку світить його робоча БС.

и) Чи змінював абонент місце роботи протягом останніх місяців.

і) Який саме пристрій використовує абонент: iPhone, Samsung та інше.

Для оцінки платоспроможності номерів інших операторів були відібрані наступні ознаки :

а) “Lifetime” - час з моменту першого дзвінка на “Київстар”. Дана характеристика слугує певною апроксимацією часу з моменту підключення.

б) Кількість дзвінків нульової довжини від абонентів “Київстар”.

в) Загальна кількість дзвінків від абонентів “Київстар”.

г) Середня тривалість дзвінків.

ґ) Середній скоринговий бал 3-х найбільш популярних номерів на які дзвонив абонент.

д) Середній скоринговий бал 3-х найбільш популярних номерів, які дзвонили абоненту.

В даній задачі дуже важливим аспектом є відслідковування динаміки в діях користувача, оскільки часто не повернення кредиту пов`язано з деякими змінами в житі людини (звільнення з роботи та інше). Для інтерпретації змін в поведінці абоненів згенеровано додаткові характериситки, як співвідношення показників за різний проміжок часу. Наприклад відношення суми взятих кредитів в “Київстар” за 3 місяці до кредитів взятих за 5 місяців.

## 3. 2 Побудова моделі

### 3.2.1 Оцінка платоспроможності абонентів “Київстар”

## 

На початку для побудови моделі для кожної характеристики була підрахована наступна метрика

m = . (3.1)

Якщо процент не нульових елементів (m) складав менше 5 %, то характеристика не враховувалась в побудові моделі.

На наступному етапі відкинуто сильно корельовані змінні. Кореляція яких по модулю складала більше 0.7.

Спочатку застосуємо декілька алгоритмів та перевіримо результати роботи моделей на 5 стратифікованих фолдах. Результати тесту занесені до таблиці 3.1.

Таблиця.3.1–Результат StratifiedKFold до бінінгу

|  |  |  |  |
| --- | --- | --- | --- |
|  | AUC (min) | AUC (avg) | AUC (max) |
| Logistic Regression | 0,57 | 0,602 | 0,632 |
| Random Forest | 0,59 | 0,625 | 0,661 |
| XGBoost | 0.636 | 0,658 | 0,687 |

Як бачимо найкращі результати показав метод XGBoost.

На наступному етапі над всіма неперервними характеристиками був проведений біннінг, для реалізації якого використовувалося дерево.

Основними характерстики побудованих дерев для бінінгу були :

* максимальна глибина дерев - 5;
* критерій оптимізації - ентропія;
* збалансованість класів;
* мінімальний процент екземплярів на кожній гілці - 20 % .

Після чого над всіма характеристиками був проведений OneHotEncoding.

На наступному етапі побудована модель логістичної регрессії зі штрафною функцією l2 (оскільки інтерес був в тому, щоб перевірити, яких якісних показників можна досягнути при використанні всього об`єму інформації) та з С = 0.001.

В результаті побудови моделі після розбиття вибірки на 10 стратифікованих фолдів було досягнуто середнє значення AUC = 0.69 на навчальній вибірці та AUC = 0.68 на тренувальній. Це доволі непогані результати, які вирішують поставлену перед нами задачу, але в моделі приймає участь велика кількість змінних (близько 800). Оскільки в українських реаліях математичне очікування цих змінних з часом може змінюватися. Наприклад середня сума затрат на оплату телеком послуг має тенденцію до постійного росту, через нестабільність в українській економіці. Тому вирішено зменшити зменшити кількість характеристик з метою покращення стійкості моделі до часових змін.

Для вибору змінних використано декількох алгоритмів.

а) RandomForest, який складався з 100 дерев з глубиною від 2 до 5 (при використанні більшої кількості характеристик алгоритм починав перенавчались.

б) XGBoost, який складався зі 100 дерев з максимальною глибиною 2 та параметрами регуляризації 𝝺 = 0.1 та 𝞪 = 1.

З двома алгоритмами відібрано 20 характеристик, які мали найбільші ваги. Обидва дали однаковий набір ключових характеристик. Після чого на цих 20 характеристиках була знову побудована модель логістичної регресії з тими самими параметрами регуляризації. В результаті чого точність моделі стала в порівняні з минулим тестом більшою. Середнє значення AUC на StratifiedKFold стало дорівнювати 0,69.

Спробуємо застосувати ще кілька алгоритмів та перевірити який показник тепер дає найбільшу точність. Результати тесту занесені до таблиці 3.2.

Таблиця. 3.2 – Результат StratifiedKFold після бінінгу

|  |  |  |  |
| --- | --- | --- | --- |
|  | AUC (min) | AUC (avg) | AUC (max) |
| Logistic Regression | 0,672 | 0,689 | 0,706 |
| Random Forest | 0,612 | 0,642 | 0,688 |
| XGBoost | 0,656 | 0,676 | 0,693 |

Як бачимо після коректної обробки даних всі алгоритми почали показувати більш високі якісні показники. Треба також відмітити, що після проведення бінінгу лідером серед алгоритмів стала логістична регрессія, яка до проведення бінінгу навпаки давала найгірші показники серед алгоритмів.

### 3.2.2 Оцінка платоспроможності абонентів інших операторів за даними компанії “Київстар”

На початку побудови моделі відкинуто приклади в яких був від’ємний показник “lifetime”, тобто вони брали кредит до першого дзвінка на “Київстар”. Оскільки по таким абонентам на момент взяття кредиту не було жодної інформації.

Також відкинуто абонентів в яких було менше 3-х подій до моменту взяття кредиту. Оскільки з їх урахуванням модель давала доволі погані показники.

Серед характеристик як й у випадку з київстаровськими абонентами були відкинуті ті, в яких процент не нульових елементів (m) складав менше 5 %.

Була побудована спочатку модель логістичної регресії, яка дала показник AUC на 10 стратифікованих фолдах 0,63. Також були побудовані моделі градієнтного бустінгу та Random Forest, які дали показники 0.65 та 0.64 відповідно.

У рішенні цієї задачі бінінг суттєво зменшив якісні показники моделей до 0,56 +- 0.02. Причиною цього є недостатній об’єм інформації про абонентів інших операторів, в таких випадках бінінг й справді є не найкращою практикою.

## 3.3 Аналіз роботи моделей

### 3.3.1 Аналіз роботи моделі для прогнозування абонентів “Київстар”

Основною задачею кредитного скорингу є ранжування клієнтів, від найбільш ризикових о найменш ризикових. Для перевірки даної метрики були використані наступні показники.

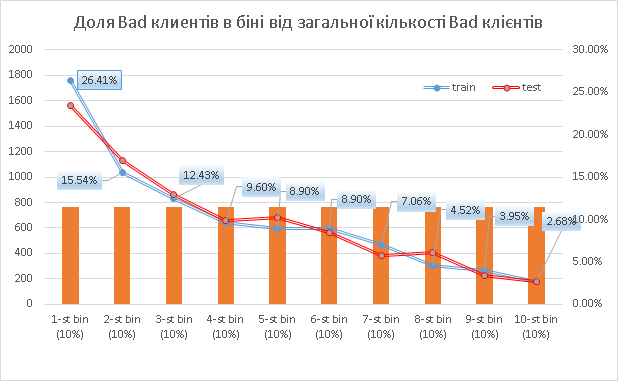
Тестова вибірка формувалася з даних за останні місяці, а навчальна з даних за попередні місяці для створення більш показового експерименту.

Якісні показники моделі занесені до таблиці 3.3.

Таблиця 3.3 – Якісні показники моделі для прогнозування платоспроможності абонентів “Київстар”

|  |  |  |
| --- | --- | --- |
|  | **Train** | **Test** |
| **AUC** | *70,1* | *71,91* |
| **GINI** | *40,2* | *41,2* |

Також навчальна та тестова вибірки були відсортовані за значенням прогнозу, а потім розбиті на 10 однакових за розмірами бінів. В кожному біні було підраховано процент поганих клієнтів від загальної кількості поганих клієнтів. Результати зображені на рисунку 3.2.

Рисунок 3.2 – Доля поганих клієнтів в біні від загальної кількості поганих клієнтів

Також в кожному біні було підраховано процент поганих клієнтів. Результат наведений на рисунку 3.3.

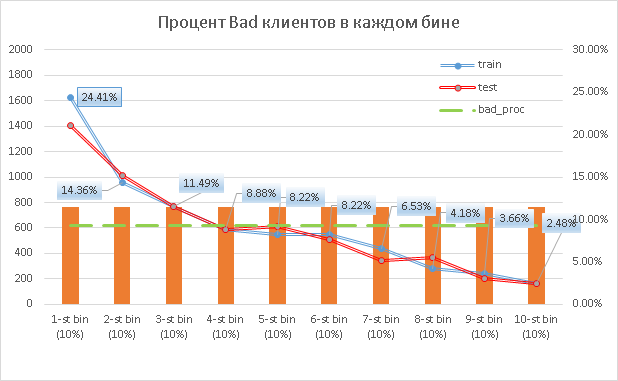


Рисунок 3.3 – Процент поганих клієнтів в кожному біні

Як бачимо з вище наведених графіків, наша модель доволі непогано ранжує клієнтів банку. З вище наведеного графіку видно, що в результаті використання нашої моделі було виявлено 30 % вибірки в якій середня кількість дефолтних клієнтів була вища в декілька разів ніж по всій вибірці. За допомогою даного графіку банк може оцінити прибуток від даної моделі та необхідне порогове значення.

Підведемо підсумок отриманих результатів та занесемо їх до загальної таблиці 3.4.

Таблиця 3.4 – Результати моделі оцінки платоспроможності абонентів “Київстар”

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Бін | Процент поганих клієнтів в одному біні | Процент поганих клієнтів в біні від загальної кількості поганих клієнтів | Середній скор в біні | Процент поганих клієнтів у вибірці |
| 1 | 24.41% | 26.41% | 0.32 | 9.32% |
| 2 | 14.36% | 15.54% | 0.22 | 9.32% |
| 3 | 11.49% | 12.43% | 0.19 | 9.32% |
| 4 | 8.88% | 9.60% | 0.183 | 9.32% |
| 5 | 8.22% | 8.90% | 0.17 | 9.32% |
| 6 | 8.22% | 8.90% | 0.159 | 9.32% |
| 7 | 6.53% | 7.06% | 0.149 | 9.32% |
| 8 | 4.18% | 4.52% | 0.139 | 9.32% |
| 9 | 3.66% | 3.95% | 0.13 | 9.32% |
| 10 | 2.48% | 2.68% | 0.115 | 9.32% |

Як бачимо перші 3 біни дають кращі результати ніж середній показник не повернення кредитів в банку. Тому рекомендованим порогом відсікання є 0,19.

### 

### 3.3.2 Аналіз роботи моделі для прогнозування платоспроможності абонентів інших операторів за даними компанії “Київстар”

Подивимося на якісні показники моделі, які занесені до таблиці 3.5.

Тестова вибірка формувалася з даних за останні місяці, а навчальна з даних за попередні місяці для створення більш показового експерименту.

Таблиця.3.5 – Якісні показники моделі для прогнозування платоспроможності абонентів інших операторів за даними компанії “Київстар”

|  |  |  |
| --- | --- | --- |
|  | **Train** | **Test** |
| **AUC** | *65,2* | *65,01* |
| **GINI** | *30,04* | *30,02* |

На наступному етапі вибірка знову була розбита на 10 бінів. Та підраховані ті самі показники, шо й для минулої моделі, результати наведені на рисунку 3.4 та рисунку 3.5.

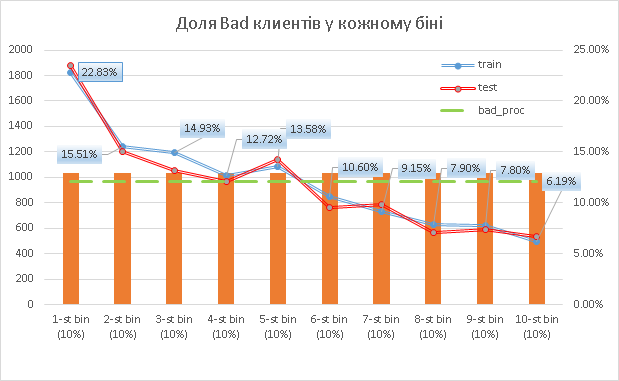


Рисунок 3.4 – Доля поганих клієнтів у кожному біні

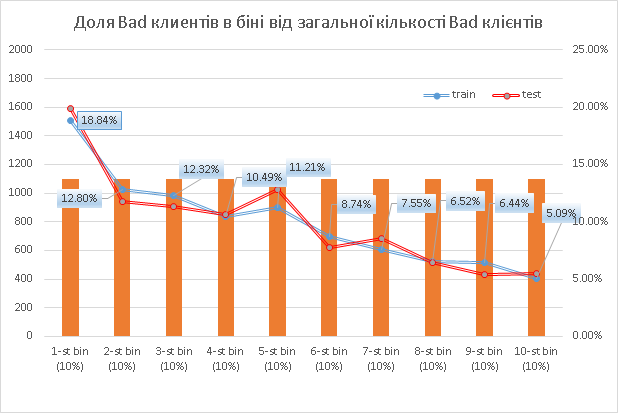


Рисунок.3.5 – Доля поганих клієнтів від загальної кількості

Занесемо всі показники до фінальної таблиці 3.5.

Таблиця 3.5 – Результати роботи моделі для оцінки платоспроможності абонентів інших операторів за даними компанії “Київстар”

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Бін | Процент поганих клієнтів в одному біні | Процент поганих клієнтів в біні від загальної кількості поганих клієнтів | Середній скор в біні | Процент поганих клієнтів у вибірці |
| 1 | 22.83% | 18.84% | 0.24002 | 12.12% |
| 2 | 15.51% | 12.80% | 0.170248 | 12.12% |
| 3 | 14.93% | 12.32% | 0.149963 | 12.12% |
| 4 | 12.72% | 10.49% | 0.137468 | 12.12% |
| 5 | 13.58% | 11.21% | 0.127672 | 12.12% |
| 6 | 10.60% | 8.74% | 0.119383 | 12.12% |
| 7 | 9.15% | 7.55% | 0.112125 | 12.12% |
| 8 | 7.90% | 6.52% | 0.104946 | 12.12% |
| 9 | 7.80% | 6.44% | 0.0973 | 12.12% |
| 10 | 6.19% | 5.09% | 0.086107 | 12.12% |

Як бачимо дана модель має гіршу роздільну здатність, але при виборі необхідного порогу відсікання банк може покращити якість свого кредитного портфелю та зекономити власні кошти.

## Висновки за розділом 3

В результаті виконаної роботи зібрані характеристики, які б могли б вплинути на платоспроможність абонентів телеком-оператора. Для побудови оцінки платоспроможності абонентів “Київстар” були обрані наступні ознаки: затрати на iнтернацiональнi дзвiнки, кiлькiсть вiдвiданих протягом останнього часу країн, час з моменту пiдключення (часто шахраї купляють новi sim-карти), поповнення рахунку, використання iнтернет трафiку, модель телефону. Складнішою виявилася задача знаходження основних характеристик для оцінки платоспроможності абонентів інших операторів за даними “Київстар”. Для вирішення даної задачі були використано такі характеристики: час з моменту першої появи абонента в мережі “Київстар”, тривалість дзвінків в мережі “Київстар”, кількість дзвінків абонентам, які знаходилися в роумінгу.

Проведена первина обробка даних. В результаті досліджень було виявлено, що при побудові моделі для оцінки платоспроможності абонентів “Київстар” суттєвих покращень в якісних показниках моделі вийшло досягти при використанні бінінгу. Найкращі показники дала логістична регресія. При побудові скорингової моделі для оцінки платоспроможності інших операторів за даними “Київстар” бінінг навпаки погіршував результати оскільки при його використані були суттєві втрати інформації. Найкращі якісні показники дала модель градієнтного бустінгу.

В результаті проведених досліджень досягнуті високі якісні показники роботи моделей, а саме AUC 0.71 на абонентах “Київстар” та 0.65 на абонентах інших операторів.

# 

# РОЗДІЛ 4 РОЗРОБКА СТАРТАПУ

## 4.1 Опис проекту

Основною метою розробленого стартапу є оцінка платоспроможності клієнтів банку по небанківським даним.

Термін реалізації проекту: 12 - 16 місяців

## 4.2 Команда проекту

Команда:

* Семен – DB administrator
* Петро – Data scientist
* Василь – Django developer
* Людовік – маркетолог
* Коля – менеджер з продажів

Задача проекту по ролям :

1. Задача DB administrator – формування вибірки для побудови прогностичних моделей, написання jobs на базу для функціонування моделей

2. Задача Data scientist – побудова прогностичних моделей та візуалізація даних.

3. Задача Django developer – загортання результатів моделей в API для зовнішнього користувача .

4 Задача маркетолог – реклами компанії

5. Задача менеджер з продажів – пошук нових партнерів та налагодження стосунків з ними

Розподіл ролей в компанії наведений на рисунку 4.1.

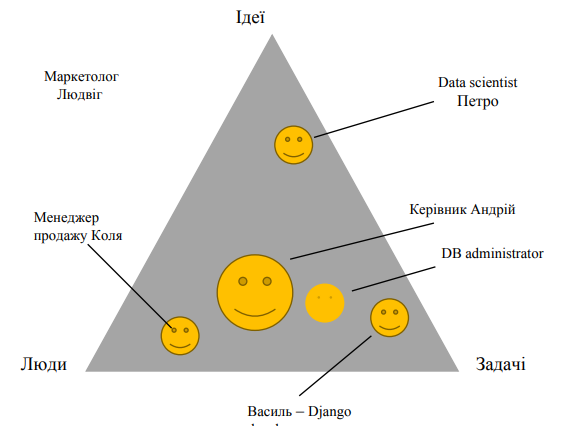


Рисунок. 4.1 – Розподіл ролей в компанії

## 

## 

## 4.3 Бізнес модель проекту

Для побудови стартапу було розроблено власний логотип, який зображений на рисунку 4.2.



Рисунок. 4.2 – Логотип стартапу

Партнери стартапу:

* телеком-оператори;
* компанії, які мають великі обсяги клієнтських даних;
* банки та фінансові установи.

Структура затрат:

* затрати на заробітню плату співробітників;
* затрати на оплату серверів ;
* виплати процентів з угод компаніям, які надають дані;
* оренда офісного приміщення ;
* апаратура для співробітників.

Структура доходів:

* оплата запросів на надання клієнтського скору.

Цінність для клієнтів:

1. Для компаній, які надають данні це в першу чергу можливість пасивного заробітку без затрат з власної сторони, також це можливість знаходження нових бізнес партнерів та піару.
2. Для банків та фінансових установ це можливість покращення власних аналітичних моделей, що в свою чергу веде за собою економію.

Сегменти користувачів:

Найбільш важливим сегментом являються компанії, які мають великі об’єми даних. Наприклад: телеком-оператори. З іншої сторони це банки, фінансові установи та інші компанії, які орієнтуються на платоспроможних клієнтів.

Причини, які призвели до розвитку даного продукту.

В першу чергу це відсутність розвинутого Data Science ринку в Україні. Більшість компаній, які обладають великими обємами даних просто не знають як вони можуть на них заробляти.

До того ж через поганий рівень ІТ спеціалістів у українській банківський сфері призводить до необхідності відання деяких задач на аутсорс. Наприклад таких, як побудова скорингових моделей, оскільки така задача для банку часто є разовою й тому їм не вигідно тримати для неї спеціалізований відділ.

Найбільш перспективні цифри:

Компанія “Білайн” отримує 80% свого прибутку на продуктах Big Data. Одним з таких продуктів є скорингові моделі для банків. Хоча звісно модель бізнесу в Росії відрізняється від Української моделі бізнесу.

Пропозиція для інвесторв:

* 60% від прибутку;
* термін реалізації від 6 до 12 місяців.

## 4.4 Аналіз ринкових можливостей стартап проекту

В таблиці 4.1 наведена попередня характеристика потенційного стартап-проекту

Таблиця 4.1 – Попередня характеристика потенційного ринку стартап-проекту

|  |  |  |
| --- | --- | --- |
| № п/п | Показники стану ринку (найменування) | Характеристика |
| 1 | Кількість головних гравців, од | 4 |
| 2 | Загальний обсяг продаж, грн/ум.од | 1000 ум.од |
| 3 | Динаміка ринку (якісна оцінка) | Зростає |
| 4 | Наявність обмежень для входу (вказати характер обмежень) | Немає |
| 5 | Специфічні вимоги до стандартизації та сертифікації | Немає |
| 6 | Середня норма рентабельності в галузі (або по ринку), % | 20% |

Основні характеристики потенційних клієнтів наведені в таблиці 4.2

Таблиця 4.2 – Характеристика потенційних клієнтів стартап-проекту

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| № п/п | Потреба, що формує ринок | Цільова аудиторія (цільові сегменти ринку) | Відмінності у поведінці різних потенційних цільових груп клієнтів | Вимоги споживачів до товару |
|  | Використання данних компанії для залучення додаткового прибутку | Телеком оператори та компанії, які мают великі ресурсі даних | Розміри оброблюваних даних, технічні обмеження | Конфеденційність наданної інформації |
|  | Покращення існуючих систем оцінки платоспроможності клієнтів | Середній та великий бізнес | Якість власних побудованих моделей | Якість прогнозування. стійкість моделей |

Основні фактори загроз наведені в таблиці 4.3.

Таблиця 4.3 – Фактори загроз

|  |  |  |  |
| --- | --- | --- | --- |
| № п/п | Фактор | Зміст загрози | Можлива реакція компанії |
| 1 | Недостатні обсяги інформації | Побудова будь-якої дата сайнс моделі є R&D здачею і ми ніколи напевно не знаємо про можливість її реалізації | 1. Попередження замовника про необхідність проведення первинного аналізу перед побудовою моделі 2. Додавання результатів роботи моделі в якості характеристики до моделі користувача, а не в якості заміни існуючої |

Фактори можливостей описані в таблиці 4.4.

Таблиця 4.4 – Фактри можливостей

|  |  |  |  |
| --- | --- | --- | --- |
| № п/п | Фактор | Зміст можливості | Можлива реакція компанії |
| 1 | Накоплення великого обєму інформації | В результаті великої кількості замовлень та постійного збору формації накопленя великих обємів даних | Створення власних дата центрів |
| 2 | Розробка продуктів таргетингової реклами | В результаті накопленя необхідної кількості інформації можлива розробка нових продуктів. наприклад таких, як таргетингова реклама. Наприклад деякий товар буде цікавий лише людям з великим рівнем заробітка, інший навпаки, в залежності від цього маркетинг компаній буде націленний не на всіх, а лише на певну аудиторію | Пропозиція нових продуктів |

В таблиці 4.5 зображений ступеневий аналіз конкуренції на ринку.

Таблиця 4.5 – Ступеневий аналіз конкуренції на ринку

|  |  |  |
| --- | --- | --- |
| Особливості конкурентного середовища | В чому проявляється дана характеристика | Вплив на діяльність підприємства (можливі дії компанії, щоб бути конкурентоспроможною) |
| 1. Вказати тип конкуренції  - олігополія | Більшисть даних належить фірмам гігантам | 1. Технологічні інновацій. 2. Кооперація з дослідницькими центрами. 3. Розширення функціоналу та задоволення потреб клієнтів. 4. Якісний набір спеціалістів |
| 2. За рівнем конкурентної боротьби  - локальний | Більшість компаній, які надають подібні сервіси розташовані на території України |
| 3. За галузевою ознакою:  -внутрішньогалузева | Продукт спрямований на роздрібну торгівлю |
| 4. Конкуренція за видами товарів:  -за бажанням | Полягає у задоволенні бажань клієнта |
| 5. За характером конкурентних переваг  - нецінова | Переваги можливі при наявності більшого обєму данних |
| 6. За інтенсивнісю  - марочна | Торгова марка впливає на довіру до компанії |

В таблиці 4.6 наведений аналіз конкуренції в галузі за М. Портером

Таблиця 4.6 – Аналіз конкуренції в галузі за М. Портером

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Скла-  дові ана-  лізу | Прямі конкуренти в галузі | Потенційні конкуренти | Пос-  тачальни-  ки | Кліє-нти | Товари-  замін-  ники |
| Бюро кредитних історій | Власні розробки великих компаній | Відсутні | Київ-  стар | Відсутні |
| Вис-  новки | Застарілі методи керівництва проектами, які не пристосовані до гнучких методів розробки. А саме ієрархічна структура та відсутність кросс-функціональності команд  Відсутність можливостей до R&D розробок | Застарілі методи керівництва проектами, які не пристосовані до гнучких методів розробки. А саме ієрархічна структура та відсутність кросс-функціональності команд  Відсутність можливостей до R&D розробок |  |  |  |

В таблиці 4.7 наведено обґрунтування факторів конкурентоспроможності.

Таблиця 4.7 – Обґрунтування факторів конкурентоспроможності

|  |  |  |
| --- | --- | --- |
| № п/п | Фактор конкурентоспроможності | Обґрунтування (наведення чинників, що роблять фактор для порівняння конкурентних проектів значущим) |
| 1 | Інновації | Інноваційні рішення мають забезпечити перевагу нашим клієнтам над конкурентами |
| 2 | Постійне накопленя інформації | Накоплення інформації з усіх можливих джерел, як і тої, яка є у відкритому доступі так і через налаштування партнерських стосунків з велиими компаніями |
| 3 | Використання методологій гнучкої розробки | Залучення кросс-функціональних команд та постійне вдосконалення продукту |
| 4 | Цінова політика | Послуги коштують дешевше, але клієтнів залучено більше |

Також був проведений зрівняльний аналіз сильних та слабких сторін проекту, який наведений в таблиці 4.8.

Таблиця 4.8 – Зрівняльний аналіз сильних та слабких сторін “DataCOM”

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| № п/п | Фактор конкурентоспроможності | Бали 1-20 | Рейтинг товарів-конкурентів у порівнянні з DataCOM | | | | | | |
| -3 | -2 | -1 | 0 | 1 | 2 | 3 |
| 1 | Інновації | 1,5 |  |  | + |  |  |  |  |
| 2 | Постійне накопленя інформації | 9 |  |  |  |  | + |  |  |
| 3 | Використання методологій гнучкої розробки | 12 |  | + |  |  |  |  |  |
| 4 | Цінова політика | 17 |  |  |  | + |  |  |  |

В таблиці 4.9 наведений SWOT-аналіз проекту.

Таблиця 4.9 – SWOT- аналіз стартап-проекту

|  |  |
| --- | --- |
| Сильні сторони:  цінова політика, використання методологій гнучкої розробки, інноваційність | Слабкі сторони:  велика кількість монополістів на дані |
| Можливості:  впровадження інноваційних рішень, розширення бізнесу, накопленя більшого обєму даних | Загрози:  нестача даних та інформації |

В таблиці 4.10 були також розглянуті альтернативи ринкового впровадження стартап-проекту.

Таблиця 4.10 – Альтернативи ринкового впровадження стартап-проекту

|  |  |  |  |
| --- | --- | --- | --- |
| № п/п | Альтернатива (орієнтовний комплекс заходів) ринкової поведінки | Ймовірність отримання ресурсів | Строки реалізації |
| 1 | Утворення стартап відділу всередині великих компаній | Середня | 6-9 місяців |

## **4.5 Розроблення ринкової стратегії проекту**

В таблиці 4.11 наведений вибір цільових груп потенційних споживачів

Таблиця 4.11 – Вибір цільових груп потенційних споживачів

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| № п/п | Опис профілю цільової групи потенційних клієнтів | Готовність споживачів сприйняти продукт | Орієнтовний попит в межах цільової групи (сегменту) | Інтенсивність конкуренції в сегменті | Простота входу у сегмент |
| 1 | Телеком-оператори | Середній | Помірний | Низька | Складний |
| 2 | Компанії, які обладають великим обємом даних | Середній | Високий | Середня | Помірний |
| 3 | Банки та фінансові установи | Високий | Середній | Висока | Не складно |

Визначення базової стратегії розвитку наведено в таблиці 4.12.

Таблиця 4.12 – Визначення базової стратегії розвитку

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| № п/п | Обрана альтернатива розвитку проекту | Стратегія охоплення ринку | Ключові конкурентоспроможні позиції відповідно до обраної альтернативи | Базова стратегія розвитку\* |
|  | Розвиток всередені великої компанії | Здобуття популярності на ринку | Аналітичні відділи,  бюро кредитних історій | Горизонтальне розширення, збільшення кількості партнерів |

Визначення базової стратегії конкурентної поведінки наведено в таблиці 4.13.

Таблиця 4.13 – Визначення базової стратегії конкурентної поведінки

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| № п/п | Чи є проект «першопрохідцем» на ринку? | Чи буде компанія шукати нових споживачів, або забирати існуючих у конкурентів? | Чи буде компанія копіювати основні характеристики товару конкурента, і які? | Стратегія конкурентної поведінки |
|  | Ні | Забирати існуючих | Буде брати краще з досвіду конкурентів | Горизонтальне розширення, збільшення кількості партнерів серед банків, фінансових установ та компаній, які мають великі обєми даних |

В таблиці 4.14 наведене визначення базової стратегії позиціонування

Таблиця 4.14 – Визначення стратегії позиціонування

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| № п/п | Вимоги до товару цільової аудиторії | Базова стратегія розвитку | Ключові конкурентоспроможні позиції власного стартап-проекту | Вибір асоціацій, які мають сформувати комплексну позицію власного проекту (три ключових) |
|  | Доступність результатів | Розвиток якості та доступності товару | Невисока вартість послуг, пошук молодих спеціалістів серед випускників вузів | Надійність, новизна, науковість |

## 4.6 Розроблення маркетингової програми стартап-проекту

В таблиці 4.15 наведено визначення ключових концепцій потенційного товару.

Таблиця 4.15 – Визначення ключових концепцій потенційного товару

|  |  |  |  |
| --- | --- | --- | --- |
| № *п/п* | Потреба | Вигода, яку пропонує товар | Ключові переваги перед конкурентами (існуючі або такі, що потрібно створити) |
| 1 | Додатковий пасивний прибуток за рахунок збережених даних | Інтелектуальну обробку збережених даних | Іноваційність, технологічна перевага |
| 2 | Великий процент неповернення кредиту | Покращення існуючих скорингових моделей | Новий ресурс даних, моделі, які будуються на достовірних даних про клієнта |

Також був проведений опис трьох рівнів моделі товару, який наведений в таблиці 4.16.

Таблиця 4.16 – Опис трьох рівнів моделі товару

|  |  |  |  |
| --- | --- | --- | --- |
| Рівні товару | Сутність та складові | | |
| І. Товар за задумом | Пасивний прибуток за рахунок даних.  Покращення банківських скорингових моделей | | |
| ІІ. Товар у реальному виконанні | Властивості/характеристики | М/Нм | Вр/Тх /Тл/Е/Ор |
| 1. Точність  2. Стійкість за часом | %  місяць |  |
| Якість: підрахунок різних статичних показників (accuracy, AUC, F score....) | | |
| Пакування | | |
| Марка: DataTelco  Товар: Make money from your data | | |
| ІІІ. Товар із підкріпленням | До продажу: прогостична модель | | |
| Після продажу: готовий продукт з інтерфейсом | | |
| За рахунок чого потенційний товар буде захищено від копіювання: за рахунок відсутності чіткої інформації про відібрані характериситики для побудови моделі | | | |

В таблиці 4.17 наведено визначення меж встановлення ціни.

Таблиця 4.17 – Визначення меж встановлення ціни

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| № п/п | Рівень цін на товари-замінники | Рівень цін на товари-аналоги | Рівень доходів цільової групи споживачів | Верхня та нижня межі встановлення ціни на товар/послугу |
|  | Ниже середнього | Високий | Середній та високий | Від 20 тис. грн |

В таблиця 4.18. наведено формування систем збуту

Таблиця 4.18 – Формування системи збуту

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| № п/п | Специфіка закупівельної поведінки цільових клієнтів | Функції збуту, які має виконувати постачальник товару | Глибина каналу збуту | Оптимальнасистема збуту |
|  | Моделі продаются 1 за одною, не можно , друга робиться лише після продажу першої | Збергання, постачання, конфеденційність | 0-1 | Надання доступу до сервера |

Концепція маркетингових комунікацій наведена в таблиці 4.19.

Таблиця 4.19 – Концепція маркетингових комунікацій

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| № п/п | Специфіка поведінки цільових клієнтів | Канали комунікацій, якими користуються цільові клієнти | Ключові позиції, обрані для позиціонування | Завдання рекламного повідомлення | Концепція рекламного звернення |
|  | Відсутність розуміння предметної області | Зутрічі, пошта | Інноваційність, економія | Залучення нових клієнтів | Можливість заробітку на своїх даних |

## 

## Висновки за розділом 4

В результаті проведеної роботи була описана ідея стартапу, її цілі, команда та сроки необхідні для реалізації даного проекту. Були детально описані основні затрати, на які підуть кошти компанії під час реалізації даного продукту та прибуток, який можна очікувати після завершення розробки.

Було досліджено стан українського ринку на сьогоднішній день. Після якого було виявлено основних партнерів та конкурентів. Були детально описані переваги та недоліки даного проекту перед конкурентами та сформована цінність, яку несе даний продукт для клієнтів.

Також була розроблена ринкова стратегія стартапу. Та можливі сценарії його подальшого розвитку. Окрім цього була сформульована маркетингова програма проекту.

# ВИСНОВКИ ЗА РОБОТОЮ ТА ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

В роботі розглянута низка питань, які пов’язані із аналізом та менеджментом кредитного ризику комерційного банку. Саме він представляє невпевненість та невизначеність у найбільш традиційній для банків України діяльності – кредитуванні.

Кредитний ризик є досить складною сукупністю елементів, які зумовлені зовнішніми (які не можуть регулюватися банком) і внутрішніми факторами. Важливою характеристикою кредитного ризику є його динамічність, вона вимагає постійного оновлення інформаційної бази про клієнтів та її збагачення. Саме тому я вирішив дослідити нову тему залучення нестандартних джерел даних для оцінювання платоспроможності клієнті, а саме дані телекомунікаційного оператора.

В якості даних для аналізу була взята актуальна вибірка з даних “Київстар”. На початку був проведений детальний аналіз цих даних, який включав в себе відкидання некоректних даних, бінімізацію неперервних характеристик, відбір найбільш впливових та стійких за часом характеристик.

На наступному етапі був проведений підбір моделей, які б давали найкращі якісні результати. Мною були застосовані такі моделі – логістична регресія, алгоритм Random Forest та алгоритм градієнтного бустінгу.

Цікавим виявився той факт, що до проведення бінімізації та відбору найкращих характеристик найкращі якісні показники давав алгоритм градієнтного бустингу, а найгірші давала логістична регресія. Але після коректної обробки даних лідером виявилася логістична регресія, яка дала AUC = 0.71, та проявила найбільшу стійкість за часом.

Для прогнозування платоспроможності абонентів інших операторів використання бінінгу навпаки погіршувало якісні показники, причиною чого була відсутність достатнього об’єму інформації про абонентів інших мобільних операторів. Найкращі якісні показники показала модель градієнтного бустінгу, AUC якої склав 0.65.

Для подальшого покращення існуючих результатів корисною для оцінювання кредитоспроможності цікавою темою для дослідження було б залучення нових джерел інформації про клієнта, а саме аналіз соціальних мереж, аналіз транзакційних даних, історії пошуку в інтернеті та інше. Також для високоякісної оцінки платоспроможності клієнта необхідне постійне динамічне оновлення вже існуючих моделей та створення нових**.** А саме дослідження можливості застосування у цій предметній області нейроних мереж та інших методів інтелектуального аналізу даних, а також комбінації різних методів.

# ПЕРЕЛІК ПОСИЛАНЬ

1. Ледок Ю. Е. Анализ кредитоспособности и оценка потенциальных заемщиков банка [Електронный ресурс] / Ю.Е.Ледок. – Режим доступа:<http://www.rusnauka.com/4._SVMN_2007/Economics/18634.doc.html>.
2. Рэдхэд К. Управление финансовыми рисками [Текст] / К. Рэдхэд, С. Хюс [пер. с англ.] – М.: НФРА-М, 1996. – 28 с.
3. Львов В.С. Финансовый анализ банков и кредитных организаций [Електронный ресурс].- Режим доступа:<http://www.cfin.ru/press/afa/97_1_022-087.pdf>
4. Ситникова Н.Ю. Кредитные риски в системе финансового риск-менеджмента: дис... канд. экон. наук: 08.00.10 / Ситникова Н.Ю. - М., 2003. – 170 с.
5. Соколова Н.В. Контроль и анализ рисков банков: дис... канд. экон. наук: 08.00.12. / Соколова Н.В. - М., 2003. – 170 с.
6. Ковальов М.C. Методика построения банковских скоринговых моделей для оценки кредитоспособности физических лиц [Електронный ресурс] / Ковальов М.C. // Интернет-журнал «Науковедение» Выпуск 2, март – апрель 2014.— Режим доступа: <http://www.bsu.by/Cache/pdf/49623.pdf>
7. Yanhao Wei - Credit Scoring with Social Network Data [Електронный ресурс] / Yanhao Wei, Pinar Yildirim, Christophe Van den Bulte, Chrysanthos Dellarocas – Режим доступу:<https://pdfs.semanticscholar.org/35fe/6cb2a49aca15341f4af6224f7411269af601.pdf>
8. Петр Руденко - Как профиль в социальных сетях позволяет банкам изучить будущего заемщика [Електронний ресурс] / Петр Руденко // Интернет-журнал Forbes. Режим доступа: [http://www.forbes.ru/finansy/internet-i-svyaz/236259-kak-profil-v-sotsialnyh-setyah-pozvolyaet-bankam-izuchit-budushcheg](http://www.forbes.ru/finansy/internet-i-svyaz/236259-kak-profil-v-sotsialnyh-setyah-pozvolyaet-bankam-izuchit-budushchego)
9. Michael A. Predicting Financial Account Delinquencies with Utility and Telecom Payment Data [Електронный ресурс] / Michael A. Turner, Ph.D. Patrick Walker, M.A. — Режим доступу: http://www.perc.net/wp-content/uploads/2015/05/Alt-Data-and-Traditional-Accounts.pdf
10. Guoping Zeng - A necessary condition for a good binning algorithm in credit scoring Applied Mathematical Sciences [Електронный ресурс] /Guoping Zeng // Режим доступу: <http://dx.doi.org/10.12988/ams.2014.44300>
11. Сорокин М.C. Построение скоринговых карт с использованием модели логистической регрессии [Електронный ресурс] / Сорокин А.C. // Интернет-журнал «Науковедение» Выпуск 2, март – март 2014.— Режим доступа: https://naukovedenie.ru/PDF/180EVN214.pdf
12. Савченко А.В. - Машинное обучение [Електронний ресурс] / Савченко А.В. // Режим доступу: https://nnov.hse.ru/data/2017/08/26/1174067834/lecture1.pdf
13. David Rosenberg - l1 and l2 regularization [Електронний ресурс] / David Rosenberg – Режим доступа: <https://davidrosenberg.github.io/ml2015/docs/2b.L1L2-regularization.pdf>
14. Mennatallah Amer - Enhancing One-class Support Vector Machines for Unsupervised Anomaly Detection [Електронний ресурс] / Mennatallah Amer, Markus Goldstein, Slim Abdennadher – Режим доступа: http://outlier-analytics.org/odd13kdd/papers/amer,goldstein,abdennadher.pdf
15. Hastie T., The Elements of Statistical Learning, / Hastie T., Tibshirani R., Friedman J.. - Stanford: Springer, 2001.- 533 p.
16. Ставицький А.В. Навчально-методичний комплекс з курсів „Прогнозування” та „Фінансове прогнозування” / А.В.Ставицький. – К.: Центр учб. лiт., 2006. – 107 с.
17. Сергей Николенко. Деревья принятия решений [Електронный ресурс]. / Сергей Николенко. — Режим доступа: <https://logic.pdmi.ras.ru/~sergey/teaching/mlcsclub/02-dectrees.pdf>
18. Milos Hauskrecht - Ensemble methods. Bagging and Boosting [Електронный ресурс]. / Milos Hauskrecht — Режим доступа: <https://people.cs.pitt.edu/~milos/courses/cs2750-Spring04/lectures/class23.pdf>
19. Trevor Hastie Bagging [Електронный ресурс]. / Trevor Hastie — Режим доступа: <http://jessica2.msri.org/attachments/10778/10778-boost.pdf>
20. Александра Д. Случайный лес [Електронный ресурс]. / Александра Д. — Режим доступа: https://alexanderdyakonov.wordpress.com/2016/11/14/%D1%81%D0%BB%D1%83%D1%87%D0%B0%D0%B9%D0%BD%D1%8B%D0%B9-%D0%BB%D0%B5%D1%81-random-forest/
21. Trevor Hastie Boosting [Електронный ресурс]. /Trevor Hastie— Режим доступа:<http://web.stanford.edu/~hastie/TALKS/boost>
22. Кузнєцова Н.В. [Порівняльний аналіз характеристик моделей оцінювання ризиків кредитування](http://www.nbuv.gov.ua/old_jrn/natural/Vkhdtu/2010_2/02_003.pdf) [Електронный ресурс] / Н.В. Кузнєцова, П.І. Бідюк // Вестник ХНТУ.-2010. - №1(34), экономика и менеджсент— Режим доступа: <http://www.nbuv.gov.ua/old_jrn/natural/Vkhdtu/2010_2/02_003.pdf>
23. Lei Tang - Cross-Valisation [Електронный ресурс] / Payam Refaeilzadeh, Lei Tang, Huan Liu – Режим доступа:http://leitang.net/papers/ency-cross-validation.pdf

# Додаток А Код програмного продукту

import pandas as pd

import calendar

import datetime

from sklearn.model\_selection import train\_test\_split

from sklearn.linear\_model import LogisticRegression

from sklearn.metrics import auc, roc\_curve

from xgboost import XGBClassifier

import numpy as np

import copy

from sklearn.tree import DecisionTreeClassifier

from sklearn.metrics import accuracy\_score

%matplotlib inline

def change\_target(data, msisdn = 'MSISDN', target = 'TARGET'):

data\_rezult = data.copy(deep = True)

num\_cnt = pd.DataFrame(data[msisdn].value\_counts()).reset\_index()

num\_cnt.rename(columns ={ 'index': 'num'}, inplace =True)

for number in list(num\_cnt[num\_cnt[msisdn] > 1]['num']):

for i in data[data[msisdn] == number].index:

data\_rezult.loc[i, target] = data[data[msisdn] == number][target].max()

return data\_rezult

def add\_months(sourcedate,months):

month = sourcedate.month - 1 + months

year = int(sourcedate.year + month / 12 )

month = month % 12 + 1

day = min(sourcedate.day,calendar.monthrange(year,month)[1])

return datetime.date(year,month,day)

def div\_tables(table1, variable1, table2, variable2, variable3, connect\_id):

merge\_tables = pd.merge(table1, table2, on = connect\_id)

rezult = pd.DataFrame(table1[connect\_id])

for i in range(len(variable1)):

rezult[variable3[i]] = merge\_tables[variable1[i]] / merge\_tables[variable2[i]]

return rezult

def generate\_categorial\_feuters(data, ident, feuters, month = 'MONTH\_DT'):

data['date'] = pd.to\_datetime(data[month], format = '%d.%m.%Y')

last\_date = data['date'].max()

feuters\_ident = copy.deepcopy(feuters)

feuters\_ident.append(ident)

data\_f = pd.get\_dummies(data[data['date'] == data['date'].max()][feuters\_ident], columns = feuters)

return data\_f.reset\_index(drop = True)

def generate\_feuters(data, ident, feuters, month = 'MONTH\_DT'):

data['date'] = pd.to\_datetime(data[month], format = '%d.%m.%Y')

last\_date = data['date'].max()

#Sum

#data\_sum\_6\_m = data.groupby(ident)[feuters].sum().reset\_index()

#data\_sum\_6\_m.rename(columns = {i:i + '\_sum\_6m' for i in feuters }, inplace = True)

data\_sum\_3\_m = data[data['date'] >= add\_months(data['date'].max(),-2)]\

.groupby(ident)[feuters].sum().reset\_index()

data\_sum\_3\_m.rename(columns = {i:i + '\_sum\_3m' for i in feuters }, inplace = True)

data\_sum\_1\_m = data[data['date'] == data['date'].max()]\

.groupby(ident)[feuters].sum().reset\_index()

#DivSum

#data\_sum\_3\_m\_6\_m = div\_tables(data\_sum\_3\_m, [i + '\_sum\_3m' for i in feuters],

# data\_sum\_6\_m,[i + '\_sum\_6m' for i in feuters]

# ,[i + '\_sum\_3m\_sum\_6m' for i in feuters],connect\_id = ident)

#data\_sum\_1\_m\_6\_m = div\_tables(data\_sum\_1\_m, feuters,data\_sum\_6\_m,[i + '\_sum\_6m' for i in feuters]

# ,[i + '\_sum\_1m\_sum\_6m' for i in feuters], connect\_id = ident)

#data\_sum\_1\_m\_3\_m = div\_tables(data\_sum\_1\_m, feuters,data\_sum\_3\_m,[i + '\_sum\_3m' for i in feuters]

# ,[i + '\_sum\_1m\_sum\_3m' for i in feuters],connect\_id = ident)

#Mean

#data\_mean\_6\_m = data.groupby(ident)[feuters].mean().reset\_index()

#data\_mean\_6\_m.rename(columns = {i:i + '\_mean\_6m' for i in feuters }, inplace = True)

#data\_mean\_3\_m = data[data['date'] >= add\_months(data['date'].max(),-2)]\

# .groupby(ident)[feuters].mean().reset\_index()

#data\_mean\_3\_m.rename(columns = {i:i + '\_mean\_3m' for i in feuters }, inplace = True)

#Max

#data\_max\_6\_m = data.groupby(ident)[feuters].max().reset\_index()

#data\_max\_6\_m.rename(columns = {i:i + '\_max\_6m' for i in feuters }, inplace = True)

#data\_max\_3\_m = data[data['date'] >= add\_months(data['date'].max(),-2)]\

# .groupby(ident)[feuters].max().reset\_index()

#data\_max\_3\_m.rename(columns = {i:i + '\_max\_3m' for i in feuters }, inplace = True)

#Merge all this bull sheet

data\_f = pd.DataFrame(data[ident].drop\_duplicates())

#data\_f = pd.merge(data\_f, data\_sum\_6\_m,how = 'left', on = ident)

data\_f = pd.merge(data\_f, data\_sum\_3\_m, how = 'left', on = ident)

data\_f = pd.merge(data\_f, data\_sum\_1\_m,how = 'left', on = ident)

#data\_f = pd.merge(data\_f, data\_sum\_3\_m\_6\_m, how = 'left', on = ident)

#data\_f = pd.merge(data\_f, data\_sum\_1\_m\_6\_m,how = 'left', on = ident)

#data\_f = pd.merge(data\_f, data\_sum\_1\_m\_3\_m, how = 'left', on = ident)

#data\_f = pd.merge(data\_f, data\_mean\_6\_m,how = 'left', on = ident)

#data\_f = pd.merge(data\_f, data\_mean\_3\_m, how = 'left', on = ident)

#data\_f = pd.merge(data\_f, data\_max\_6\_m,how = 'left', on = ident)

#data\_f = pd.merge(data\_f, data\_max\_3\_m,how = 'left', on = ident)

return data\_f

def correlation(data, threshold):

dataset = data.copy(deep = True)

col\_corr = set() # Set of all the names of deleted columns

corr\_matrix = dataset.sample(10000).corr()

for i in range(len(corr\_matrix.columns)):

for j in range(i):

if corr\_matrix.iloc[i, j] >= threshold:

colname = corr\_matrix.columns[i] # getting the name of column

col\_corr.add(colname)

if colname in dataset.columns:

del dataset[colname] # deleting the column from the dataset

return dataset

def binning(dataset, target, columns):

data = dataset.copy(deep = True)

X\_train, X\_test, y\_train, y\_test = train\_test\_split(data[columns],

data[target].as\_matrix()

, test\_size=0.2, random\_state=2010)

# binning

l = []

ranges = {}

no\_bin = [i for i in data.columns if i not in columns]

for col in columns:

ratio = np.count\_nonzero(data[data[col] > 0][col]) / len(data)

if ratio >= 0.05:

tree = DecisionTreeClassifier(criterion='entropy', max\_depth=5,

min\_samples\_leaf=0.18,

class\_weight='balanced',

random\_state=255)

curr\_column = np.swapaxes(np.array([(X\_train[col].values)]), 0, 1)

tree.fit(curr\_column, y\_train)

classes = tree.tree\_.apply(np.swapaxes(np.array([(data[col].values)]), 0, 1).astype('float32'))

if min(classes) != max(classes):

data[col + '\_BIN'] = classes

l.append([col, len(set(classes)), 'good'])

for clazz in set(classes):

ranges[col + '\_BIN\_' + str(clazz)] = [min(data[data[col + '\_BIN'] == clazz][col]),

max(data[data[col + '\_BIN'] == clazz][col])]

else:

l.append([col, 'bad'])

data.drop(col, axis=1, inplace=True)

for cf in [i for i in data.columns if i not in no\_bin] :

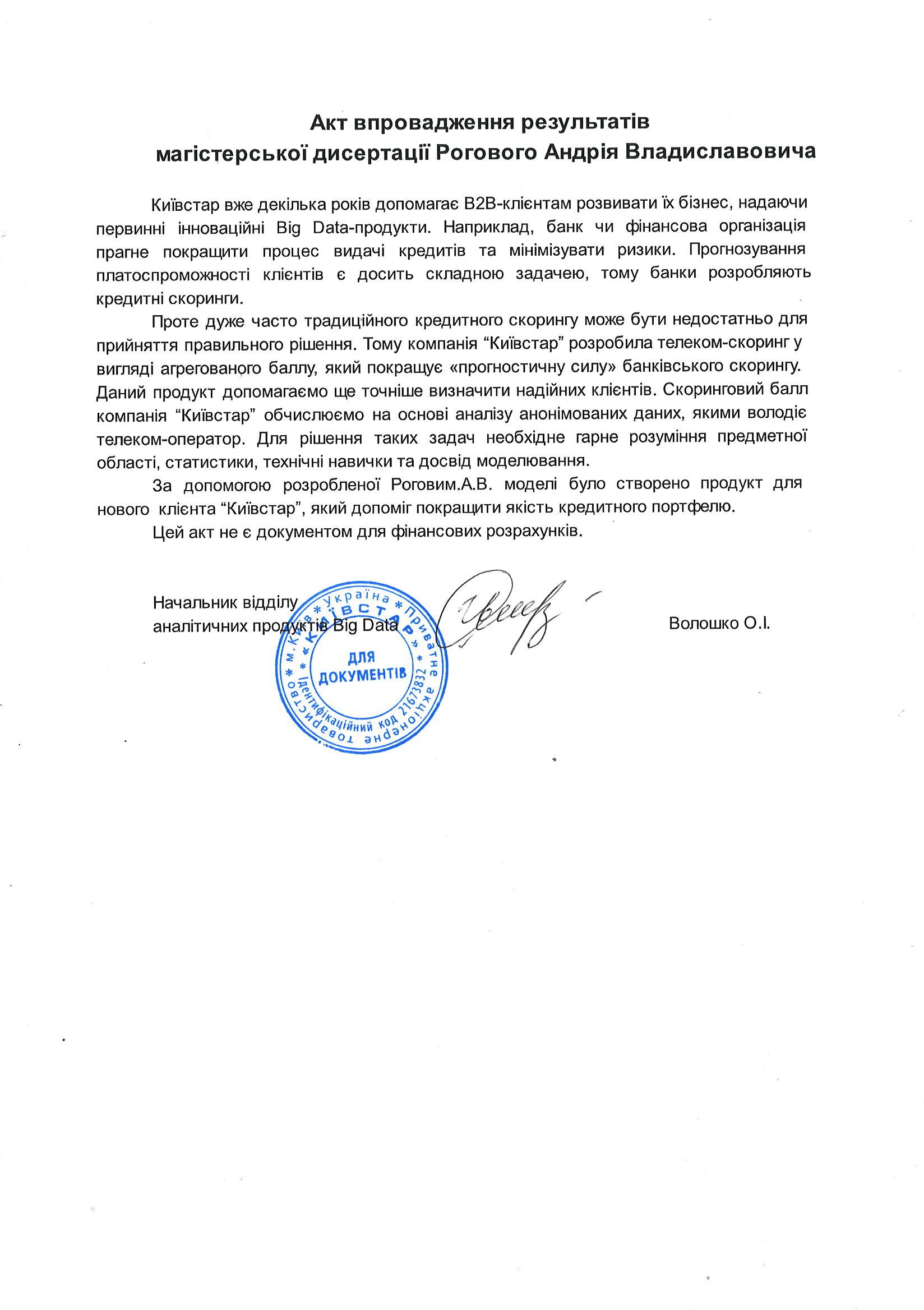
dummy\_f = pd.get\_dummies(data[cf], prefix=cf, prefix\_sep='\_')

data = pd.concat([data, dummy\_f], axis=1)

data.drop(cf, axis=1, inplace=True)

return data, ranges

# Додаток Б Акт впровадження



# Додаток В Ілюстративні матеріали доповіді

