**КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ**

**ІМЕНІ ТАРАСА ШЕВЧЕНКА**

**Факультет інформаційних технологій**

Кафедра технологій управління

**на тему: «Вдосконалення методів прогнозування для покращення кредитної системи в Україні»**

Київ

2019

**ЗАВДАННЯ**

**на магістерську роботу**

1. Виконати дослідження проблеми підвищення якості скорингового аналізу даних клієнтів-позичальників кредитів фінансових установ.

2. Перелік завдань, які потрібно розробити: 1. Зібрати необхідні статистичні дані і вибрати методи моделювання для оцінювання кредитоспроможності клієнтів. 2. Побудувати моделі для прогнозу платоспроможності. 3. Виконати аналіз отриманих результатів

3. План аналізу проблем магістерскої роботи: 1. Актуальність теми методи та мета дослідження. 2. Постановка задачі. 3. Джерела кредитної інформації 4. Що можна дізнатися з даних про клієнта ? 5. Препроцессінг даних. 6. Logistic regression. 7. Decission tree. 8. Ансамблі дерев. 9. Побудова моделей. 10. Результати роботи моделі. 11. Висновки.

Календарний план

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| № | Назва етапів виконання | Термін виконання | Примітка |  |
| з/п | дипломної роботи | етапів роботи |  |
|  |  |
|  |  |  |  |  |
| 1 | Отримання завдання | 01.12.2018 |  |  |
|  | Збір інформації | 15.12.2018 |  |  |
| 2 | Дослідження | 28.01.2019 |  |  |
|  | існуючих рішень. |  |  |  |
| 3 | Дослідження алгоритмів прогнозування | 10.03.2019 |  |  |
|  |  |  |  |  |
| 4 | Перевірка точності прогнозу | 15.03.2019 |  |  |
|  | методів алгоритмів |  |  |  |
| 5 | Розробка гіперпараметрів | 25.03.2019 |  |  |
|  |  |  |  |  |
| 6 | Виконання | 25.04.2019 |  |  |
|  | алгоритмів |  |  |  |
| 7 | Перевірка точності обраних алгоритмів | 30.04.2019 |  |  |
|  | з підбором гіперпараметрів |  |  |  |
| 8 | Оформлення дипломної роботи | 31.05.2019 |  |  |
|  | Виконання алгоритмів | 25.04.2019 |  |  |
| 9 | Передзахист роботи | .05.2019 |  |  |
| 10 | Захист роботи | 31.05.2019 |  |

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? На щастя, розвиток технологій не стоїть на місці, застосуємо нові методи прогнозування.

**Мета роботи**: розробка моделей для оцінювання платоспроможності клієнтів Укріїни.

**Об’єкт дослідження**: дані українського банку АТ «Альфа-Банк» на основі яких виконувалося прогнозування кредитоспроможності клієнтів.

**Предмет дослідження**: математичні моделі та методи обробки статистичних даних, критерії якості даних.

**Методи дослідження**: логістична регресія, градієнтний бустинг, Random forest (англ. випадковий ліс).

**Наукова новизна одержаних результатів**: наразі існує кілька методів прогнозу для побудови банківської моделі:

- Логістична регресія;

- Дерева прийняття рішень(CART).

Дані методи провірені часом, проте технології вдосконалюються і нині існує нові методи побудови, які показують кращу точність моделі статистичної моделі, - логістична регресія(з покращеними характеристиками), градієнтний бустинг, Random forest (англ. випадковий ліс).

**Практичне значення одержаних результатів**: 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 банківських моделей(скорингових моделей), у зв’язку з цим банки постійно шукають нові нестандартні джерела кредитування. Застосування нових методів моделювання – є новою задачею, яка потребує для розв`язання окрім необхідних знань, також генерування нових ідей та «підходів».

# РОЗДІЛ 1 ОЦІНКА ПЛАТОСПРОМОЖНОСТІ КЛІЄНТА

## 1.1 Сутність кредитного ризику і методи боротьби з ним.

Кредитний ризик (англ. credit risk) – наявний або потенційний ризик для надходжень і капіталу, який виникає через неспроможність сторони, що взяла на себе зобов'язання, виконати умови будь-якої фінансової угоди із банком 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:

Демографічні показники: вік, стать, національність, місце проживання, тривалість проживання в актуальному місці, освіта, професія, тривалість працевлаштування, наявність власності, сімейний стан, наявність дітей і інше.

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.

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.

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 Аналіз існуючих у світі способів її вирішення.

## Скорингова модель - це методика оцінки кредитного ризику, яка дозволяє, оцінивши набір ознак, що характеризують позичальника, визначити, чи варто надавати йому кредит.

## В основі скорингової системи лежить припущення, що люди зі схожими соціальними показниками поводяться однаково. Суть скорингу полягає в тому, що кожному параметру, що характеризує позичальника, надається реальна оцінка в балах.

## Побудова скорингової моделі проходить такі етапи:

## **C:\Users\User\Desktop\практика\Схема.PNG**

Основними методами побудови скорингової моделі серез банків України:

- Логістична регресія;

- Дерева прийняття рішень(CART).

1.3 Визначення вхідних даних для вирішення поставленої задачі.

Для прикладу була використана тестова вибірка даних. За початкові дані були взяті репрезентативна вибірка банку АТ «Альфа-Банк».

Вибірка складається із 12 змінних:

1. Стан поточного рахунку;
2. Кредитна історія:
   * відсутні кредити;
   * всі кредити виплачені належним чином;
   * існуючі кредити повернуті належним чином дотепер;
   * затримка в окупності в минулому;
   * наявність інших наявних кредитів (не в цьому банку).

3. Мета кредиту

* автомобіль
* меблі / обладнання
* радіо / телебачення
* побутова техніка
* ремонт
* освіти
* відпустка
* перепідготовка
* бізнесу
* інші

1. Сума кредиту
2. Ощадний рахунок / облігації
3. Стаж поточної роботи
4. Нерухомість

нерухомість

якщо не A121: договір заощаджень будівельних компаній /

страхування життя

якщо не A121 / A122: автомобіль або інше, не в атрибуті 6

невідомо / немає власності

1. Ставка внеску у відсотках від наявного доходу
2. Вік