НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ ІГОРЯ СІКОРСЬКОГО»
ІНСТИТУТ ПРИКЛАДНОГО СИСТЕМНОГО АНАЛІЗУ
КАФЕДРА МАТЕМАТИЧНИХ МЕТОДІВ СИСТЕМНОГО АНАЛІЗУ

На правах рукопису
УДК 004.855.5:519.876.2

До захисту допущено
В. о. завідувача кафедри ММСА
Тимощук О. Л.
«___»_____2020 р.

Магістерська дисертація

на здобуття ступеня магістра за спеціальністю 122 Комп'ютерні науки
на тему: «Оцінювання платоспроможності методами машинного навчання»

Виконав:
студент ІІ курсу, групи КА-93мп
Скидан Богдан Олегович _____

Керівник:
доцент кафедри ММСА,
к.т.н., доц. Тимошенко Ю.О. _____

Рецензент:
доцент кафедри прикладної математики
КПІ ім. Ігоря Сікорського,
к.т.н., доц. Маслянко П.П. _____

Засвідчую, що в цій магістерській дисертації
немає запозичень із праць інших авторів
без відповідних посилань
Студент _____

Київ
2020

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 — 122 «Комп'ютерні науки»

ЗАТВЕРДЖУЮ
В. о. завідувача кафедри ММСА
О. Л. Тимощук
«___» _____ 2020 р.

ЗАВДАННЯ
на магістерську дисертацію студента Скидана Богдана Олеговича

1. Тема дисертації: «Оцінювання платоспроможності методами машинного навчання», науковий керівник дисертації Тимошенко Юрій Олександрович, доцент кафедри ММСА, к.т.н., затверджені наказом по університету № 3182-с від «02» листопада 2020;

2. Термін подання студентом дисертації: 14 грудня 2020 р;

3. Об'єкт дослідження: платоспроможність фізичних осіб;

4.Предмет дослідження: моделі оцінювання платоспроможності та ймовірності дефолту фізичних осіб-позичальників;

5. Перелік завдань, які потрібно розробити:

1) дослідити сучасний стан та особливості методів оцінювання платоспроможності;
2) провести огляд сучасних підходів для побудови моделей оцінки платоспроможності;
3) обрати та обґрунтувати вибір декількох моделей
4) обробка вхідних даних;
5) застосувати оброблені дані на реалізованих моделях та проаналізувати отримані результати;
6) розробити стартап-проект виведення на ринок результатів дослідження;
7) розробити концептуальні висновки за результатами наукового дослідження.

6. Орієнтовний перелік графічного (ілюстративного) матеріалу:

1. Рисунки схем роботи деяких алгоритмів машинного навчання;
2. Таблиці порівняння результатів роботи;

3. Рисунки та графіки на яких зображено результати роботи;
4. Таблиці у розділі стартап-проекту.

7. Орієнтовний перелік публікацій:

1) Тимошенко Ю.О., Скидан Б.О., Оцінювання платоспроможності методами машинного навчання, *X Міжнародна науково-практична онлайн конференція «MODERN SCIENCE: PROBLEMS AND INNOVATIONS»*, 13-15 грудня 2020р. , Стокгольм, Швеція. С. 201-203 URL:https://sci-conf.com.ua/wp-content/uploads/2020/12/MODERN-SCIENCE-PROBLEMS-AND-INNOVATIONS-13-15.12.20.pdf

8. Дата видачі завдання: 1 вересня 2020 р.

Календарний план

з/п	Назва етапів виконання магістерської дисертації	Термін виконання етапів магістерської дисертації
1	Концептуальний вступ дисертації. Формулювання об'єкта, предмета, цілі, завдань, новизни, практичної значущості результатів	05.09.2020—13.09. 2020
2	Перший розділ. Огляд літературно-інформаційних джерел, формування нормативної бази. Характеристика об'єкта.	16.09.2020—27.09.2020
3	Другий розділ. Математичні розрахунки та сегментація показників оцінювання платоспроможності. Застосування математичних моделей для прогнозування дефолту позичальника.	30.09.2020—18.10.2020
4	Третій розділ. Обробка вхідних даних. Реалізація та застосування моделей. Збір та аналіз результатів.	21.10.2020—29.10.2020
5	Четвертий розділ. Стартап-проект	30.10.2020—17.11.2020
6	Концептуальні висновки. Перспективи розвитку отриманих рішень	22.11.2020—26.11.2020

Студент Скидан Б.О.
Науковий керівник дисертації Тимошенко Ю.О.

РЕФЕРАТ

Магістерська дисертація: 78 с., 24 табл., 36 рис., 1 дод., 31 джерел.

МАШИННЕ НАВЧАННЯ, ВИПАДКОВИЙ ЛІС, ЛОГІСТИЧНА РЕГРЕСІЯ, ОЦІНКА ПЛАТОСПРОМОЖНОСТІ, МЕТОД ОПОРНИХ ВЕКТОРІВ

Метою цього дослідження є вдосконалення існуючої методології оцінювання платоспроможності фізичних осіб, проведення моделювання методами машинного навчання, збір результатів та порівняння основних характеристик різних систем машинного навчання.

Об’єктом дослідження є платоспроможність фізичних осіб та моделі машинного навчання.

Предметом дослідження виступає база даних транзакцій по кредитам фізичних осіб та моделі машинного навчання для оцінки кредитної платоспроможності клієнтів.

В роботі описано загальні методи для оцінки кредитної платоспроможності, які використовують різні фінансові установи. Також наведено методи машинного навчання, які було застосовано для проведення моделювання, загальні принципи роботи з даними, метрики для коректного аналізу результатів роботи моделей.

Наукова новизна полягає у автоматизації прийняття кредитних рішень за допомогою різних методів машинного навчання та порівняння результатів для виявлення найбільш придатних моделей для розв’язання поставленої задачі.

ABSTRACT

The master's thesis:78 p., 24 tables, 36 fig., 1 add., 30 sources.

MACHINE LEARNING, RANDOM FOREST, LOGISTIC REGRESSION, SOLVENCY FORECAST, SUPPORT VECTOR MACHINE

The aim of this work are development of current methods for solvency forecasting of individuals, conduct modeling using machine learning methods, achieve results and complaining of main characteristics for different machine learning systems, and present the results.

The object of the research is the solvency of individuals and different machine learning methods.

The subject of the research is a database of transactions on loans to individuals and models for evaluation the creditworthiness of customers.

The research describes general methods for creditworthiness forecasting used by various financial institutions. Also it consists methods of machine learning used for current modeling, general principles of data science, metrics for correct analysis of modeling results.

The scientific novelty consists in the automation of credit decision-making by different methods of machine learning and comparison of results to choose the most suitable models for solving the current problem.

ЗМІСТ

ВСТУП

Актуальність проблеми. Оцінка платоспроможності фізичних осіб є важливим завданням для більшості фінансових установ в сучасних умовах. Також зростання попиту на такі послуги пов’язане зі значним розвитком апаратного забезпечення, алгоритмів інтелектуального аналізу даних та багатьох суміжних областей. Тож, сьогодні маючи декілька співробітників спеціалізованих на Data Science, кредитні заклади можуть розраховувати на значне покращення якості оцінювання платоспроможності, передбачення дефолту позичальників та разом із цим збільшити і власний прибуток або мінімізувати втрати.

Важливою частиною науки про дані є алгоритми та методи машинного навчання. Саме вони, як інструмент аналізу кредитної історії клієнта є одним із тих ключових засобів, що дозволяють отримати конкурентну перевагу на ринку кредитування. Основна задача що постає перед банками — це бінарна класифікація, тобто моделям потрібно маючи деяку інформацію про клієнтів відділити надійних позичальників від ненадійних. Завдяки розвитку апаратного забезпечення, став можливий розвиток алгоритмів машинного навчання, збільшення їх швидкості роботи, покращення точності результату, значно зріс обсяг інформації доступний для навчання таких моделей. Ще однією перевагою таких систем це відсутність так званого «людського фактору».

Оскільки ця методика є повністю незалежна від впливу людини вона мінімізує ризики, що пов’язані з помилками персоналу.

Але залишається відкритим питання, які саме моделі краще використовувати. Відповіді на це питання і присвячена дана дисертація.

РОЗДІЛ 1. ДОСЛІДЖЕННЯ ПРЕДМЕТНОЇ ОБЛАСТІ

Платоспроможність у сучасному житті грає велику роль, як для фінансових установ так і для звичайних людей. Для звичайних людей це поняття означає можливості якими вони можуть користуватися, блага які їм доступні й рівень життя на який вони можуть очікувати.

Для установ це вже зовсім інший аспект їх роботи. Якісний аналіз кредитоспроможності клієнтів це необхідність для різних фінансових компаній. Вони потребують сучасних, якісних та швидких методів для розв’язку задачі кредитного оцінювання. І чим кращі будуть такі системи, тим на більші прибутки можуть очікувати власники.

Отже фінансові установи найбільш зацікавлені в розвитку методів оцінки платоспроможності позичальників. Оскільки це надає їм ринкову перевагу перед конкурентами.

1.1 Поняття кредитоспроможності

Основою кредитних відносин між позичальником та клієнтом є кредитна платоспроможність. Саме вона є основним критерієм для погодження видачі кредиту фінансовою установою. І її оцінка повинна бути чітко регульована самою установою.

Але для правильної оцінки потрібно навести чітке визначення.

Платоспроможність — здатність позичальника обслуговувати борг, тобто виплачувати борг та проценти до нього в строки зазначені в кредитному договорі.[1]

Отже, як можна побачити з визначення це можливість позичальника виконувати свої зобов’язання.

Основною проблемою для кредитних установ є чітке визначення розміру та строків кредиту.

Як показано нижче на рисунку 1.1, при вирішенні цієї проблеми, звертають увагу на декілька груп ознак:

1) банківська історія клієнта або наскільки вчасно сплачувались внески по минулим кредитам;
2) кількість коштів від продажу активів позичальника в разі відмови сплати внесків;
3) джерела надходження грошей до клієнта на сьогодні.

Рисунок 1.1 – Складові кредитоспроможності

Але є один аспект, що не дозволяє повністю автоматизувати видачу кредитів. Це елемент довіри. Оскільки не існує загальноприйнятих та обов'язкових методів, методик чи норм оцінки платоспроможності, кожна установа сама визначає для себе можливі способи перевірки інформації про клієнта, оцінки його платоспроможності та умови кредитного договору.[2,3]

Отже трапляються ситуації, коли декільком позичальникам із приблизно однаковою фінансовою ситуацію працівники установ призначають кардинально різний рівень кредитоспроможності.

Тобто в кожному окремому випадку організація сама визначає який кредитний ризик брати на себе.

1.2 Огляд існуючих методів оцінки платоспроможності

Як показано на рисунку 1.2 існує дві великі групи методів оцінки платоспроможності [3-5]:

1) класифікаційні (статичні);
2) комплексного аналізу.

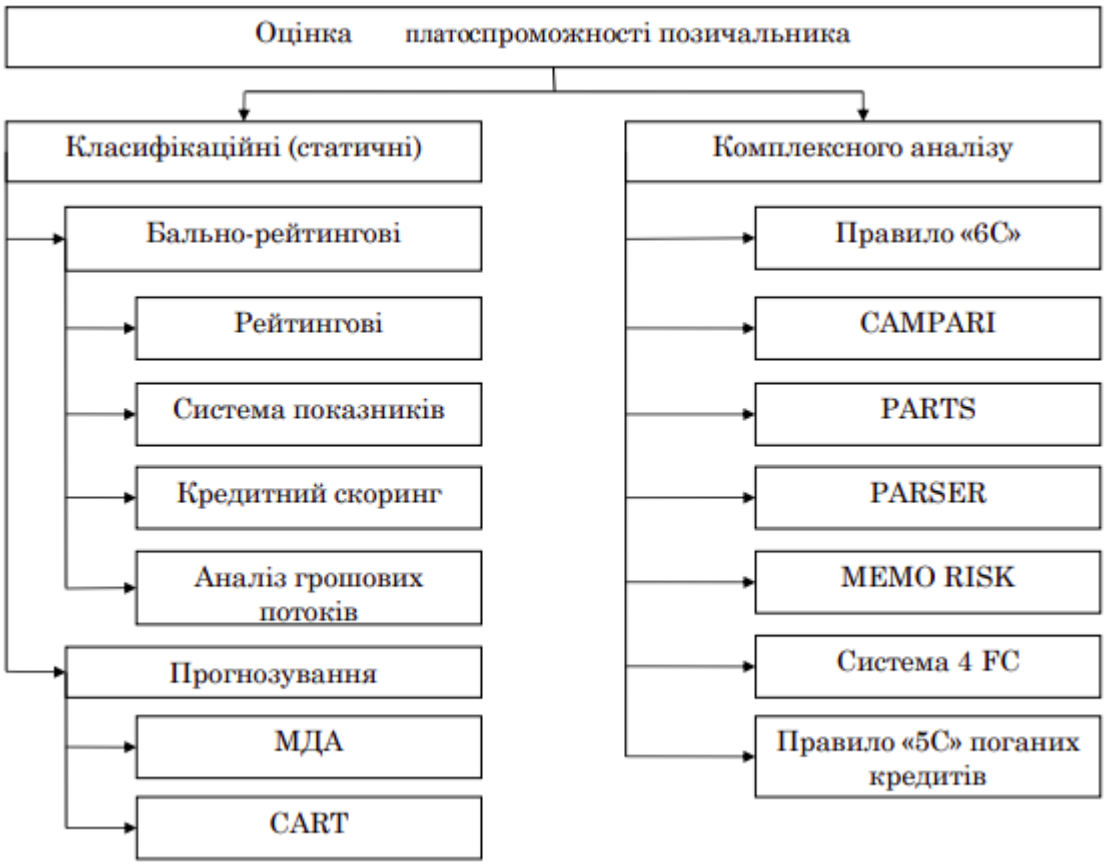

Рисунок 1.2 – Методики оцінки платоспроможності

Класифікаційні або статичні методи передбачають перевірку стандартними методами на основі матеріалів що надали самі позичальники та подальший поділ їх на дві категорії: надійних та ненадійних. Серед переваг таких методів це розповсюдженість та формалізованість. Але є й недоліки, до них належать такі фактор як обсяг інформації з яким потрібно працювати.

Також тут не враховуються неформальні ознаки, які складно точно описати та охарактеризувати.

Бально-рейтингова система дозволяє оцінити своєчасність майбутніх платежів, вартість та суму власності клієнта, оцінити його фінансові можливості та їх періодичність, а також дає змогу визначити межі зменшення обсягів доходу, у якому погашається частка постійних платежів. Позитивними сторонами таких моделей є легкість у реалізації, можливість розрахувати підходящі значення для окремих показників, можливість сортувати клієнтів за метриками, системний підхід до оцінки платоспроможності (використовуються ознаки, що показують декілька аспектів фінансового стану клієнта).

Розглянемо моделі комплексного аналізу. У таблиці 1.1 наведені декілька представників та описані значення.

Таблиця 1.1 — Розшифровка деяких назв методів комплексного аналізу платоспроможності

CAMPARI (Деякі європейські банки)	PARSER (Великобританія)	6C (США)	PARTS
C—character; A — ability; M— marge; P — purpose; A — amount; R — repayment; I — insurance.	P — person; A — amount; R — repayment; S — security; E — expediency; R — remuneration.	C — capacity; C — character; C — capital; C — collateral; C — conditions; C — confidence.	P — purpose; A — amount; R — repayment; T — term; S — security.

Далі розібрано детально їх особливості. Вони відрізняються від попередніх тим що вони використовують складні системи ознак, більшість із яких встановлюють оцінюють незалежні експерти. Поширені ці моделі зазвичай у розвинених країнах Європи та США. Відомим їх представниками є модель 6C [6].

6C — це модель оцінювання платоспроможності в якій особливу увагу приділяють таким аспектам:

1. фінансові можливості (capacity to pay) — банк хоче знаки як і коли ви будете повертати кошти перед тим як погодити вашу позику. Аналізуються ваші фінансові операції протягом деякого проміжку часу — зазвичай 2–3 роки. Історія платежів за попередніми позиками аналізується з погляду своєчасності їх оплати. Також враховуються альтернативні джерела погашення — це додаткові джерела доходу, які можна використати для повернення боргу. Сюди можуть входити особисті активи, заощадження або рахунки в банках, які можуть бути використані;
2. особистість (character) — це дуже суб'єктивна оцінка позичальник. Кредиторам потрібно вірити, що клієнт — це надійна людина, на яку можна покластися, яка поверне позику. Деякі професійні характеристики, такі як особиста кредитна історія, кваліфікація та досвід роботи, — це деякі лише деякі фактори які розглядають на цьому етапі. Банки хочуть вести бізнес із чесними і справедливими людьми. (Різниця між можливістю та бажанням повернути позику є прикладом характеру людини);
3. капітал (capital) — особистий капітал клієнта. Клієнти що мають вклади в різні компанії й готові ризикувати ним позитивно виглядають в очах кредитодавців, оскільки такі особи мають досвід у веденні бізнесу. Також гарно впливає на рішення кредиторів і наявність власного бізнесу;
4. застава (collateral) — вважаються машини, дебіторська заборгованість, запаси та інші господарські активи, які можна продати, якщо позичальник не може сплатити позику. Також важливо додати що такі дрібні речі, як комп'ютери та офісне обладнання, як правило, не вважаються заставою, у випадку більшості довгострокових позик для фізичних осіб для затвердження кредиту потрібні особисті активи власника (наприклад, його будинок чи автомобіль). Коли позичальник

використовує свої особисті активи як гарантію під ділову позику, це означає, що установа може продати ці особисті активи для задоволення будь-якої непогашеної суми, яка не була повернена;

5. умови (conditions) — це загальна оцінка економічного клімату. Економічні умови, характерні для мети позичальника, що претендує на позику, а також загальний стан економічного фактору країни значною мірою впливають на рішення про затвердження позики. Очевидно що в процвітаючих економіках більш розвинена банківська справа загалом, а зокрема кредитна її частина. Мета позики є важливим фактором. Типові фактори, включені в цей етап оцінки, включають: розмір, привабливість ринку та будь-які відповідні соціальні, економічні й політичні сили, які можуть вплинути на шанс повернення позики;
6. довіра (Confidence) — позичальник вселяє довіру до позикодавця, розглядаючи всі проблеми кредитора щодо інших п’яти С. Їх заявка на позику повідомляє, що компанія є професійною, має чесну репутацію, хорошу кредитну історію, обґрунтовану фінансову звітність, хорошу капіталізацію та належне забезпечення.

Подаючи заявку на позику, не треба забувати про важливість особистих стосунків. Спробуйте отримати позику в банку, де у вас уже склалися позитивні ділові стосунки. Крім того, зробіть спробу зустрітися з особою, яка буде розглядати вашу заявку, наприклад, з банківським службовцем, а не з касиром, який обробляє ваші повсякденні банківські операції.

Також доцільно навести модель PARTS [7], комплексного аналізу яка дуже поширена закордоном, розшифровується абревіатура як:

p (purpose) — ціль, призначення, мета позички. Експертами аналізується мета кредиту та можливі умови того що кредит не буде повернуто;

a (amount) — кількість, сума, розмір кредиту. Тут аналізується яку кількість доцільно брати в позику для обраної мети;

r (repayments) — прибуток кредиторів або відсотки. На цьому етапі обирають ціну кредиту, тобто яку суму прибутку отримає установа;

t (Term) — строк, на який надається кредит, період часу за який позичальник повинен повернути борг;

s (Security) — безпека, перестраховка, забезпечення кредиту. Тобто тут розглядаються поручителі, застави й будь-які інші джерела резервного доходу у випадку банкрутства позичальника.

Отже ця модель оцінювання платоспроможності спирається менше на особистість і більше на сам кредит, тут не має суб'єктивних ознак, усе чітко описується і ясно обраховується.

Ще одна широко розповсюджена закордоном модель комплексного аналізу[8], це CAMPARI:

c (character) — репутація, особистість, суб'єктивна оцінка того хто бере позичку;

a (ability) — можливість, потенціал клієнта до повернення боргу вчасно;

m (marge) — націнка, накрутка, прибуток, процент організації;

p (purpose) — обґрунтування мети для якої береться позичка;

a (amount) — аналіз суми кредиту, чи дійсно для заданої мети потрібні такі кошти;

r (repayment) — на цьому етапі обґрунтовують умови погашення, строк;

i (insurance) — розглядаються поручителі, застави й будь-які інші джерела резервного доходу у випадку банкрутства позичальника.

Як можна помітити ця модель деякими аспектами перетинається з попередньою, але є більш суб'єктивною, тут застосовується експертна оцінка особистості можливого клієнта.

Метод PARSER дозволяє поетапно підходити до аналізу шести областей інтересу. Що важливо, він визначає мету позики.[8]

Розшифрування PARSER:

p (person) — особистість позичальника, його здобутки;

a (amount) — аналіз суми, чи потрібні для такої мети такі кошти;

r (repayment) — можливість, умови строк виплати заборгованості;

s (security) — забезпечення, тобто можливі активи якими буде перекриватися борг у разі несплати позики;

e (expediency) — задумка, мета, користь, аргументація причини взяття позики;

r (remuneration) — банківський процент (дохід установи).

При цій методиці більш виділяється особистий аспект клієнта. Характеристики позичальника аналізуються з культурно-етичної точки зору. Основними ознаками для розгляду є рішучість клієнта погасити борг.[9]

Методика 4FC (4 Foundations of Creditworthiness — 4 ознаки кредитної платоспроможності) оцінює насамперед якість менеджменту, що аналізується фінансовими можливостями та професійністю можливого боржника; стан економічної області, яка окреслюється мережевими й мінливими обставинами в галузі, а також ринковим станом замовника; можливістю продажу застави, яка перевіряється вартістю активів які були записані як застава по кредиту; можливі фінансові передумови, які ґрунтуються на дослідженні прибутковості та ліквідності за певний проміжок часу.[10,11]

Ще одна вагома методика це MEMO RISK:

management — управління, тобто суб’єктивна експертна оцінка управління позичальника;

experience — досвід роботи особи, що хоче взяти кредит;

market — загальні ринкові умови для ведення діяльності позичальника;

operations — загальна оцінка всієї діяльності можливого клієнта;

repayment — опис можливих та ймовірних ситуацій які можуть вплинути на ймовірність повернення кредиту позичальником;

interest — банківський процент, тобто відсоткова ставка яка являється прибутком кредитора;

security — усі можливі активи, кошти, застави, поручителі, перші внески, які збільшують ймовірність того що установа не понесе збитки працюючи з даним клієнтом.

Як бачимо ця методика здебільшого зациклена на прибутку установи, та майже повністю виключає експертну складову оцінки платоспроможності.

Якісна оцінка платоспроможності позичальників також надає перевагу не тільки провести аналіз фінансового стану господарства або фізичних осіб, але ще й помітити проблемні аспекти в їх господарському менеджменті та дає можливість покращити майбутній вектор розвитку підприємства. Такі послуги надані вчасно, відкривають перспективи щодо змін кредитних відносин позичальників та кредитно-фінансових установ.

1.3 Огляд результатів досліджень

Важливим етапом дослідження предметної області є аналіз вже існуючих результатів у цій сфері.

Значущою виявилась стаття в якій описано і порівняно класичні методи машинного навчання для задачі оцінювання кредитного оцінювання для фізичних осіб. [12]

Там порівнювались такі класичні методи ІАД, як:

1) LDA (linear discriminant analysis) - лінійний дискримінантний аналіз;
2) logit Analysis - логіт аналіз або логістична регресія;
3) k-nearest Neighbors - метод k найближчих сусідів;
4) classification and regression Trees (CART) - класифікація методами регресійних дерев (дерев рішень);
5) neural networks - нейронні мережі.

Автори цієї статті наголошували про те що більшість тогочасних фінансових установ використовували як основний математичний метод аналізу кредитної інформації саме логістичну регресію. Це пов'язано з тим що для роботи логістичної регресії у більшості випадків не треба робити ніяких додаткових припущень. Інші методи, такі як CART або нейронні мережі,

використовувались в основному як допоміжні засоби, або в процесі вибору змінних чи в процесі оцінки якості моделі. Менш відомі kNN метод взагалі не застосовувався, а якщо і застосовувався то дуже нечасто.

На той момент отримані результати показали що нові й більш прогресивні методи мали чудовий потенціал і були конкурентоспроможними в задачах класифікації, порівняно з логістичною регресією.

Автори зазначали що у більшості випадків на практиці вибір методу був залежний не від його ефективності а від досвіду осіб які займались цим питанням. Також важливим аспектом ефективності різних моделей на практиці є кількість та якість даних якими володіють фінансові установи. Об'єм вибірки, якість ознак, метод збору інформації є важливим аспектом ефективності кредитного аналізу.

Також було сказано що, непараметричні методи, на відміну від логістичної регресії, можуть успішно працювати з відсутніми значеннями та мультиколінеарністю (або кореляцією) серед змінних, але часто є дуже вимогливими до апаратного забезпечення. Результати, отримані на основі деяких з цих методів, важко пояснити менеджерам, а також клієнтам. Наприклад, незважаючи на те, що нейронні мережі можуть показувати чудові результати, неможливість пояснити, причину може бути серйозним недоліком застосування цих методів в практиці.

Іншою статтею, було дослідження пов'язане з аналізом цифрового сліду як доповнення до класичних методів кредитного аналізу.[13]

Там показано що використовуючи навіть прості, доступні змінні, які можна зібрати в онлайн сервісах фінансових установ та застосувавши нескладні моделі машинного навчання, можна отримати результат який значно перевищить по якості кожен з підходів окремо. Це говорить про те, що цифрова інформація доповнює, швидше ніж замінники інформації кредитних установ, який використовує інформацію з обох джерел (кредитна оцінка + цифровий слід) може приймати кращі рішення щодо кредитування порівняно з кредиторами, які мають доступ одне з двох джерел інформації.

Автори зазначають, що важливою перепоною до повного розповсюдження таких моделей є те що цифрові сліди можна змінювати, використовуючи певне програмне забезпечення. Але як додатковий інструмент для оцінки платоспроможності такий спосіб оцінювання клієнтів є досить перспективним.

Ще одним важливим дослідженням стала стаття в американському журналі. В якій було досліджені аспекти кредитного оцінювання в умовах Big Data (великих даних).[14]

Автори прийшли до того що спосіб оцінювання позичальників для видачі кредитів повинен бути юридично сформульованим, обґрунтованим та прозорим. Щоб кожна особа яка отримала відмову, могла чітко розуміти чому. Це пов’язано з тим що кредитна галузь є важливою частиною суспільного процвітання в багатьох країнах. Отже використання деяких моделей, таких як метод опорних векторів, які не мають чітких пов’язаних між собою критеріїв оцінки, а лише видають найбільш ймовірний результат є неприпустимим. Також у висновках до цього дослідження сказано про важливість юридичного врегулювання таких процесів та неможливість сліпому слідування результатам моделей.

1.4 Висновки до розділу 1

Кредитна платоспроможності є важливою частиною всієї банківської справи. Це один із найбільших джерел прибутку для таких установ. І її якісна оцінка являє собою запоруку успіху для кредитних закладів. Без якісного розуміння кредитної платоспроможності неможлива їх робота.

У цьому розділі було наведено визначення платоспроможності, розкрито її складові, показана актуальність обраної теми. Показано основні аспекти кредитного відбору. Також частину розділу було присвячено опису

загальновідомих та поширених у світовій практиці методик оцінки платоспроможності. Були описані такі методики як PARSER, CAMPARI, 4FC, PARTS, MEMO RISK, 6C та розглянуто особливості їх використання.

Важливою частиною цього розділу є огляд результатів минулих досліджень. Описано результати 3 досліджень, які значним чином вплинули на сьогоднішній стан в області застосування методів машинного навчання для фінансових послуг.

РОЗДІЛ 2 МЕТОДИ МАШИННОГО НАВЧАННЯ ДЛЯ ОЦІНЮВАННЯ ПЛАТОСПРОМОЖНОСТІ

Сьогодні методи машинного навчання відіграють ключову роль у багатьох сферах пов’язаних з інформаційними технологіями.

У сучасному світі пройшов різкий якісний розвиток інформаційних технологій. Для роботи в нових умовах значних змін зазнали й алгоритми роботи з даними. Бурхливий прогрес, не лише машинного навчання, а й усієї сфери штучного інтелекту став можливий завдяки технологізації суспільства, яка в свою чергу призвела до формування великих вибірок даних та покращення якості апаратного забезпечення. Це забезпечило стрімкий розвиток алгоритмів машинного навчання, що дозволило покращити їх якість.

Насамперед, машинне навчання покликане надавати максимально точні результати, базуючись на вже зібраних даних, щоб замовники, маркетологи і користувачі могли якісно оцінювати інформацію якою вони володіють, і на її основі приймати рішення.

Основним елементом успіху машинного навчання є накопичення досвіду. Чим більше даних було використано під час навчання тим кращий результат можна отримати.

У цьому розділі описано основні наукові аспекти дослідження, показано проблеми з якими автор зіштовхнувся та способи їх розв’язання.

Машинне навчання є сукупністю методів для роботи з даними. Воно моделює навчання за принципом навчання людини. Так званого методу «спроб і помилок» і чим більше цих спроб і помилок тим краще вийде результат. Навчання проводиться за допомогою деяких даних зібраних людьми. Зазвичай, для коректної роботи алгоритмів і досягнення потрібної якості результату, потребується значний обсяг інформації.

Останніми роками машинне навчання грає значну роль в автоматизації оцінювання платоспроможності. У цьому розділі буде наведено та описано методи та моделі за допомогою яких було проведено дане дослідження.

2.1 Балансування вибірки

Основним завданням моделей потрібних для розв'язання поставленої проблеми є бінарна класифікація. Ми будемо визначати до якого класу належить клієнт, до того що повертає кредит чи ні. На основі вхідних та обчислених даних кредитну історію позичальника та про сам кредит, класифікатори прогнозують ймовірність позитивного (або негативного) результату із ймовірністю:

$$P(positive) = 1 - P(Negative)$$

Однак в нашому випадку кількість повернених та неповернених кредитів дуже сильно відрізняється. Наближено на 1 "поганий" кредит знайдеться 20 "хороших". Якщо ж у такому випадку всі кредити позначити як "хороші", то точність прогнозу досягне приблизно 95%, але при цьому такий прогноз не буде мати ніякої цінності.

Традиційні методи машинного навчання розраховані насамперед на збалансовані набори даних. Найпопулярніші алгоритми класифікації, такі як метод опорних векторів (SVM), нейронні мережі та дерева рішень спрямовані на оптимізацію своїх цільових функцій, які зазвичай призводять до максимальної загальної точності - відношення кількості правдивих прогнозів до всіх прогнозів.[15-16]

Коли ці методи навчаються на дуже незбалансованих наборах даних, вони часто обирають клас-більшість. Коли більшість вибірки даних складає клас з позитивним результатом, ці методи матимуть високий рівень істинних передбачень позитивних прикладів(True Positive Rate), але низький рівень правильних передбачень коли приклад є негативним (True Negative Rate). Деякі дослідження показали що для стандартних класифікаторів збалансований набір даних забезпечує покращення загальної ефективності класифікації порівняно з незбалансованим набором даних [17].

На рисунку 2.1 показано, що існує декілька способів боротьби з незбалансованістю даних, а саме undersampling та oversampling, тобто додавання або зменшення кількості прикладів у вибірці та генерація синтетичних даних.[18]

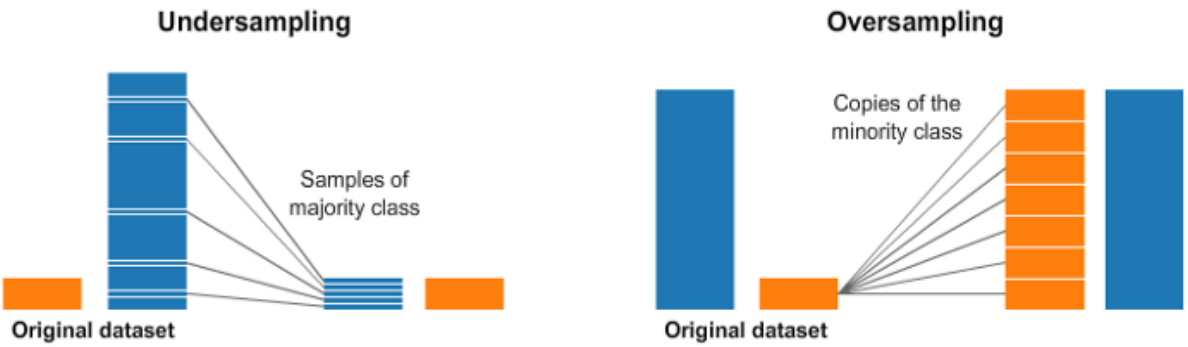

Рисунок 2.1 – Принципи роботи методів балансування вибірки

Undersampling - метод балансування вибірки який на меті має зменшити кількість прикладів з класом більшістю(majority). Цей метод застосовують якщо вибірка є достатньо великою. При видаленні даних з вибірки модель може пропустити досить важливі ознаки притаманні класу більшості, що в свою чергу викликає збільшення кількості помилок першого роду, коли кредити з високим шансом повернення будуть віднесені до “неблагополучної” категорії. Основною перевагою цього методу є те що, на великих вибірках даних при його застосуванні в рази зменшується час навчання моделей. [19]

З іншого боку метод oversampling не видаляє дані а випадковим чином дублює вже існуючі дані з класом меншістю (minority). І тут вже виникає інша проблема, а саме проблема перенавчання оскільки ідентичні приклади формуються у вибірці.[20]

У випадку додавання прикладів ймовірно ми зіштовхнемося з проблемою перенавчання. Тобто класифікатор може нормально працювати лише на даній вибірці, а на реальних даних прогнози будуть досить низької якості. Також важливою перепоною є те що складність багатьох алгоритмів машинного навчання є функцією квадратичною або навіть й більшої степені від кількості прикладів. Збільшення елементів вибірки може призвести до багатократного збільшення часу навчання моделі.

З першого погляду здається що ці методи є функціонально ідентичними, але наслідки, які з’являються після використання кожного з них є різними.

Також важливим методом є класифікація з урахуванням витрат (cost-sensitive). Він передбачає те що не всі помилки класифікації є ідентичними. Наприклад при видачі позики “поганому” клієнтові банк втрачає свої кошти. Якщо класифікатор неблагополучного клієнта розпізнає як благополучного, це призведе до втрати ресурсів. В ситуації коли модель не видасть кредит “хорошому” клієнтові, то банк не втрачає нічого. Отже такі методи працюють з метою зменшення кількості загальних втрат, коли одні з них важливіші за інші. На рисунку 2.2 показано схематично, як модель з урахуванням витрат буде розпізнавати різні приклади у вибірці даних.[21]

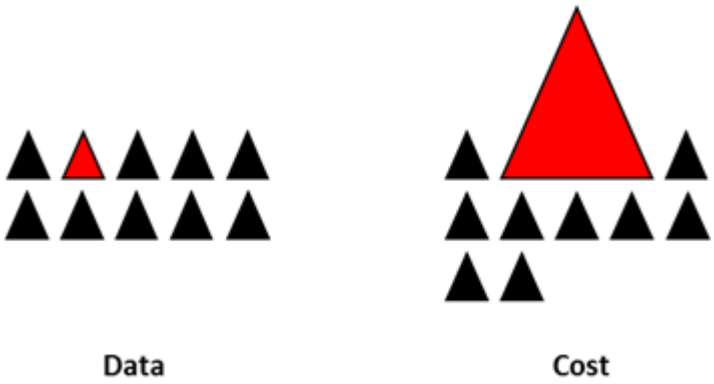

Рисунок 2.2 – Схема роботи алгоритмів з урахуванням витрат

Так у дослідженні було обрано метод з додаванням прикладів з “поганими” кредитами.

2.2 Логістична регресія

Першу модель яка була застосована на даних була логістична регресія.

Логістична регресія - це статистична модель, яка базується на логістичній функції для прогнозування бінарно залежної змінної, хоча існує багато більш складних розширень з іншими функціями та для більш складних змінних.

Логістична регресія виконує класифікації з двома можливими результатами. Тут передбачається саме ймовірність приналежності до одного класів. Це модифікація моделі лінійної регресії для задач класифікації. [22,23]

Використовується логістична регресія у базовому варіанті для випадку моделювання системи з 2 класами класифікації. Система моделює ймовірність того, що приклад вибірки відповідає одному з двох можливих значень. Але потрібно пам’ятати, що модель передбачає лише ймовірність, а не клас. Вибір класу залежить від обраного порогу передбачення.

Перевагою логістичної регресії також є те що можна задати ваги для кожної окремої змінної. Ці коефіцієнти визначають вплив кожної ознаки на результат моделювання.

Також логістичну регресію потрібно використовувати коли множина класів у вибірці є лінійно розподільною: Логістична регресія проводить плавну, лінійну межу прийняття рішення між двома класами. Таким чином, якщо ваші класи лінійно розділяються, логістична регресія буде працювати дуже добре.

Основною функцією у лінійній регресії, яка оцінює ймовірність приналежності об’єктів до класів є сигмоїда (Sigmoid). Графік її наведено на рисунку 2.3. Та описується вона як функція логістичного розподілу :

$$\Theta(z) = \frac{1}{1 + e^{-z}}$$

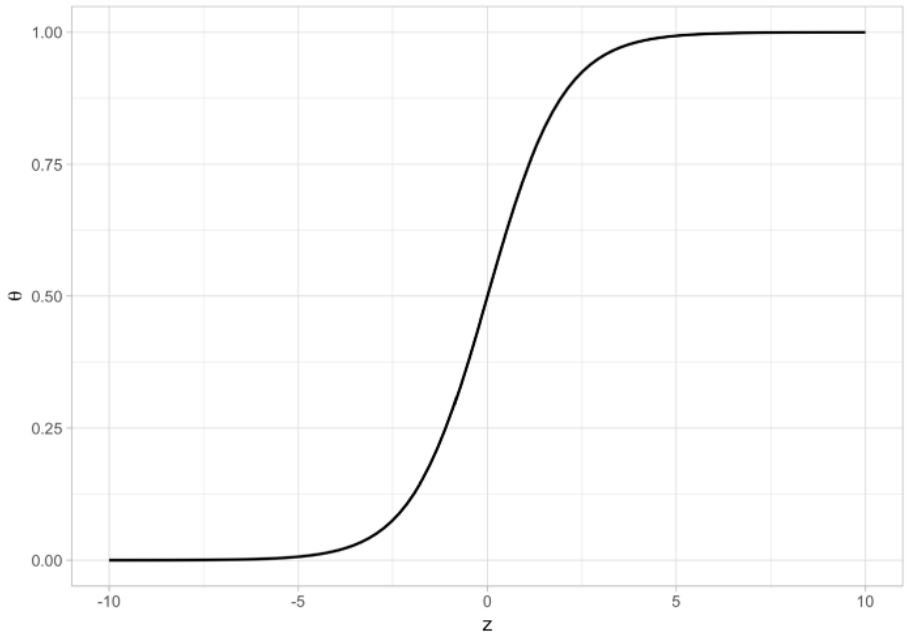

Рисунок 2.3 – Графік сигмоїдної функції

Перевагою такої функції є її додатні значення та властивість обмеженості на всій множині дійсних чисел. Вона знаходиться в межах 0 та 1

на всій множині значень. Що робить її ідеальним варіантом для відтворення ймовірності.

Також потрібно зазначити що при балансуванні вибірки логістична регресія може працювати значно краще. [24]

Логістичну регресію можна зрозуміти просто як обчислення β параметрів, які найкраще підходить для розв'язання системи:

$$y = \begin{cases} 1, \text{якщо } \beta_1 + \ \beta_2 x + \ \varepsilon > 0 \\ 0 \text{ , у всіх інших випадках} \end{cases},$$

де ε - помилка розподілена за логістичним розподілом,а x - навчальний приклад.

2.3 Метод опорних векторів

Наступним методом який було використано у роботі став метод опорних векторів (Support Vector Machine — SVM)

"Support Vector Machine" (SVM) - це керований алгоритм машинного навчання, який загалом використовують для класифікації, але може бути застосований і для регресії. В алгоритмі будується кожен елемент вибірки даних, як точку в n-мірному просторі, де n - кількість ознак, які містяться у тренувальному наборі, кожен з елементів якого має власну координату у цьому просторі. Потім проводиться класифікація, знаходячи гіперплощину, яка дуже добре розділяє n-вимірний простір на два класи, як показано на рисунку 2.4.

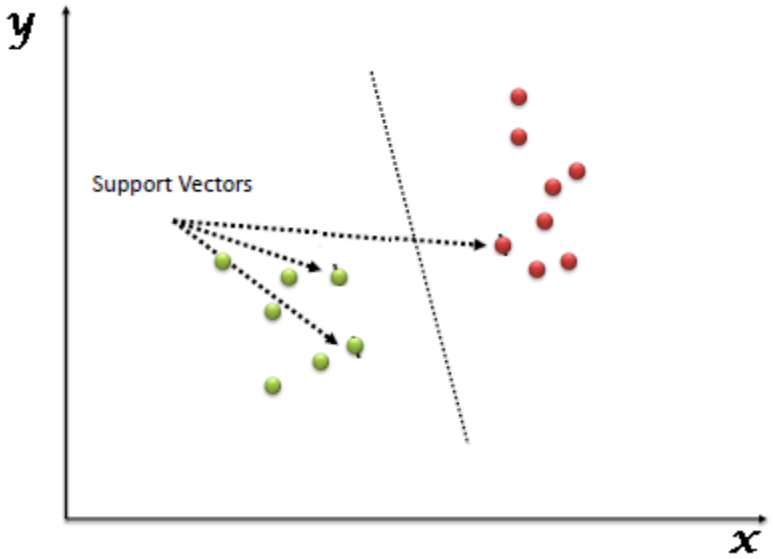

Рисунок 2.4 – Приклад роботи SVM.

Метод опорних векторів будує деяку гіперплощину або їх набір у просторі тренувальних даних, які можуть бути використані для класифікації або виявлення викидів. Слід зазначити що розмірність таких гіперплощин буде складати n-1, де n - кількість об’єктів у навчальній вибірці.[25]

Якісної класифікації досягає гіперплощина, яка має найбільшу різницю квадратів координат точок навчальних даних одного класу, оскільки загалом чим далі знаходяться точки від гіперплощини, тим краще працює модель. Це продемонстровано на рисунку 2.5.

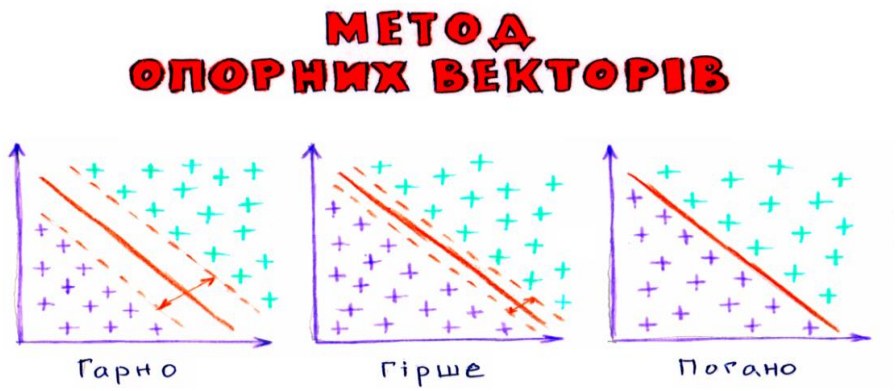

Рисунок 2.5 – Критерій якості методу опорних векторів

Тепер запишемо формальну задачу методу опорних векторів:

Є набір тренувальних прикладів (x_1,y_1) .... $,(x_n,y_n)$ коли $y_1,\dots y_n$ приймає значення -1 якщо точка не належить до класу та 1 якщо належить. Треба побудувати таку площину щоб відстань від неї до будь-якої точки x_1 була максимальною.

Площину можна записати як

$$\underline{w} * \underline{x}\text{-}b = 0,$$

де $\underline{w}$ є вектором нормалі до площини;

b зсув відносно початку координат.

Приклад такої площини показано на рисунку 2.6.[21]

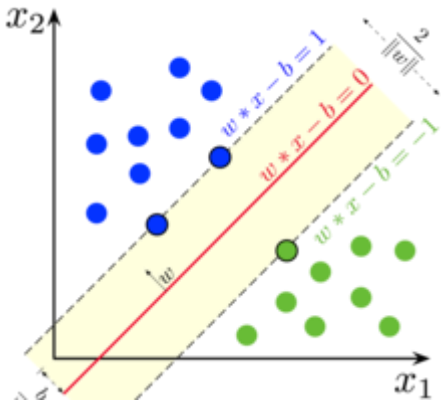

Рисунок 2.6 – Приклад розділової площини натренованої на двокласовій вибірці

Якщо вибірку можна лінійно розподілити, то є можливість побудувати дві паралельні площини, які поділять простір на дві частини які будуть відповідати двом класам, таким чином що максимальна різниця між довжинами векторів є якомога більшою. Область, яка обмежена площинами,

називається поділом , а максимально роздільна площина є гіперплощиною, яка знаходиться посередині між ними. Вони описуються системою рівняннь:

$$\begin{cases} \bar{w} * \bar{x} - b = 1 \\ \bar{w} * \bar{x} - b = -1 \end{cases}$$

З математичної точки зору, відстань L між двома площинами визначається як:

$$L = \frac{2}{\left|\left|\underline{w}\right|\right|};$$

Тоді для максимізації її між ними потрібно мінімізувати $\left|\left|\underline{w}\right|\right|$, отже маємо дві нерівності:

$$\begin{cases} \bar{w}\bar{x} - b \geq 1, \text{якщо } y_i = 1 \\ \bar{w}\bar{x} - b \leq -1, \text{якщо } y_i = -1 \end{cases}$$

Запишемо їх в одну:

$$y_i(\bar{w}\bar{x}_i\text{-}b), \text{ для всіх } 1 \leq i \leq n$$

Отже сформулюємо задачу оптимізації повністю:

Мінімізувати $\left|\left|\underline{w}\right|\right|$ за умови $y_i(\bar{w}\bar{x}_i\text{-}b)$, для всіх i = 1,...,n

Очевидним є те що, така площина буде визначатися тими $\bar{x}_i$ що знаходяться найближче до неї. Такі $\bar{x}_i$ й будуть називатись опорними векторами.

Важливим покращенням алгоритму опорних векторів є заміна скалярного добутку на нелінійну ядрову функцію, що дозволило поширити даний метод на вибірки з лінійно нерозподіленими класами, яких більшість у сучасних умовах.[26]

Найбільш поширеними функціями ядра методу опорних векторів є:

1) Поліноміальна:

$$k(x_i, x_j) = (x_i \cdot x_j + 1)^d ;$$

2) Гаусова радіально-базисна функція:

$$k(x_i, x_j) = exp\left(-\gamma \left|\left|x_i - x_j\right|\right|^2\right);$$

3) Сигмоїдна (функція гіперболічного тангенсу):

$$k(x_i, x_j) = \tanh(kx_i \cdot x_j + c) \text{ для } k > 0, c > 0.$$

Ефективність методу опорних векторів сильно залежить від підбору параметрів. Часто такий підбір проводиться так званими методами “грубої сили” (англ. - brute force). Він представляє собою звичайний перебір всіх скінченних комбінацій параметрів.

Тобто беруть кожен можливий варіант параметрів, запускають його на існуючих даних, і обираються ті параметри що показують найкращу точність.

Найчастіше обирають функцію ядра Гаусову радіально-базисну функцію оскільки вона має лише один параметр γ. Також підбирають коефіцієнт м'якого розподілу C.

Як можна побачити на рисунку 2.7, беруться попарно всі можливі значення параметрів C та γ які зростають експоненціально :

$$C \in \{2^{-5}, 2^{-3}, \ldots, 2^{13}, 2^{15}\}$$

$$\gamma \in \{2^{-15}, 2^{-13}, \ldots, 2^{1}, 2^{3}\}$$

Рисунок 2.7 – Значення параметрів C та γ

Оскільки кількість таких комбінацій складає n^2,де n це кількість можливих значень у вибірці, то моделі потрібно буде відпрацьовувати n^2кількість разів, через це оптимізація методу займає багато часу та обчислювальних потужностей комп'ютера.

Ще одним недоліком є неможливості інтерпретувати коефіцієнти натренованої моделі, тобто коефіцієнти моделі не відображають впливу окремих змінних на результат.

2.4 Випадковий ліс (Random forest)

Наступним алгоритмом застосованим у дослідженні став Random forest. Це метод який використовує побудову дерев рішень на етапі тренування і видає клас який є модою класів (для класифікації) або середнє значення класів (для регресії). Алгоритм є дуже схильним до перенавчання, тому важливо правильно підбирати тренувальні вибірки.[27]

Для порівняння Random forest, зазвичай, краще працюють ніж звичайні дерева рішень, але їх точність нижча, ніж у алгоритму gradient boosted trees. Однак властивості вибірки даних можуть вплинути на ефективність цього алгоритму.[26]

Основним принципом роботи алгоритму є так звана “мудрість натовпу”, тобто той факт що декілька некорельованих моделей будуть працювати краще ніж кожна окремий їх представник. Тобто узагальнення отриманого досвіду. У бінарній класифікації така модель буде працювати за принципом “більшості”. Приклад наведено на рисунку 2.8: 6 дерев передбачило значення прикладу як 1, та 3 дерева передбачило 0 – отримуємо значення 1.

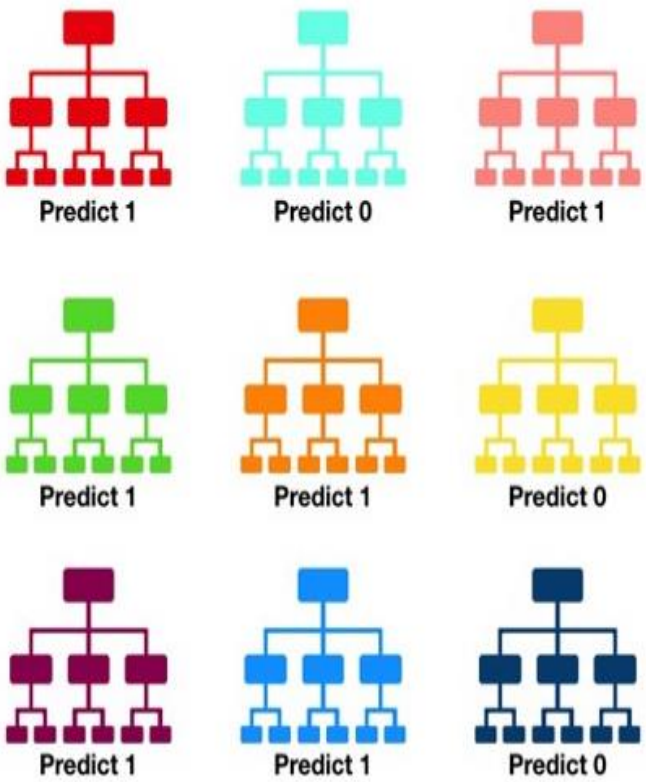

Рисунок 2.8 - Ідея алгоритму Random forest

Random forest є ансамблевим алгоритмом класифікації і при його роботі використовуються два основні ідеї:

1) метод випадкових підпросторів;
2) метод беггінга (усереднення результатів).

Цей алгоритм часто використовують як "чорний ящик" в реальних проектах, оскільки він генерує досить точні прогнози для досить широкого спектру даних, досить легкий у дослідженні, та не потребує спеціалістів з значним досвідом роботи у сфері інтелектуального аналізу даних.

Формальний запис алгоритму:

Маємо тренувальну вибірку з N-елементів$X = \underline{X}_1, \ldots, \underline{X}_N$ та відповіді $Y = Y_1, \ldots, Y_N$

Генеруємо B, випадкових підвибірок X, Y розміром n<N таким чином щоб приблизно N/3 елементів не потрапило до жодної підвибірки та будуємо дерево рішень на цих підвибірках.

Ті елементи вибірки $\hat{x}$ які не потрапили до тренувальних підвибірок генерують своє значення як середнє значення прогнозів по іншим підвибіркам.

$$f = \frac{1}{B}(\sum_{b=1}^{B} f_b(\hat{x})$$

Оптимальну кількість дерев B можна знайти, використовуючи перехресну перевірку, або спостерігаючи середню помилку прогнозування для кожної навчальної вибірки x_i , використовуючи лише дерева, які не мали x_i у своїй вибірці.

Потрібно зазначити, що перевагою алгоритму є стійкість до шумів та викидів. Це зумовлено тим що кожне окреме дерево може бути вразливим для шумів, але після застосування процедури беггінга, тобто усереднення результатів роботи множини дерев, вони стають невразливими до таких помилок.

Крім того, невизначеність прогнозу σ може бути обрахована як стандартне відхилення передбачень від усіх окремих дерев на x':

$$\sigma = \sqrt{\frac{\sum_{b-1}^{B}(f_b(\mathrm{x}') - \widehat{f})^2}{B-1}},$$

де f_b -значення певного дерева b;

$\hat{f}$ –середнє значення по всім деревам.

Кількість дерев B, є вільним параметром. Зазвичай використовується від декількох сотень до декількох тисяч дерев, залежно від розміру та характеру навчальної вибірки.

Помилки в тренувальній та тестовій вибірці, як показано на рисунку 2.9, зазвичай , стають сталими після певної кількості тренувальних епох.

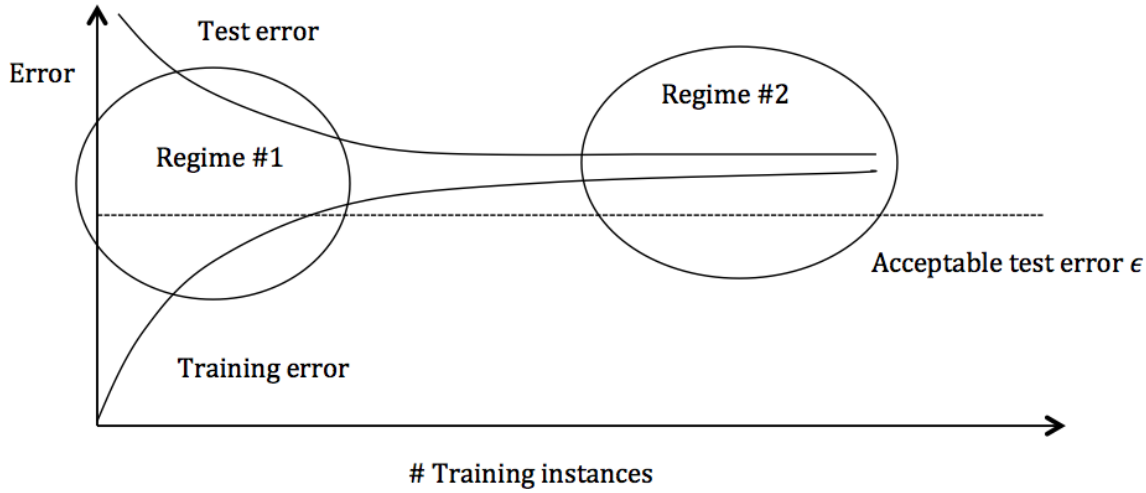

Рисунок 2.9 – Рівень помилок на тестовій та тренувальній вибірках залежно від кількості тренувальних епох.

2.5 Метрики для оцінювання якості моделей класифікації

Після моделювання дуже важливим етапом роботи дослідника є аналіз зібраних результатів. В цьому дуже допомагають існуючі метрики.

Першою метрикою яка використовується для оцінки якості класифікації є точність прогнозу (accuracy) і описується вона за формулою:

$$acc = \left(\frac{TRUE}{ALL}\right) * 100\%$$

де True - кількість правильно прогнозованих елементів вибірки;

All = кількість всіх елементів вибірки.

Ця метрика є простою, зрозумілою і широко використовуваною, але є в неї і один значний недолік. Тут ми нічого не можемо сказати про тип помилок, які допустив класифікатор.[27]

Як видно на рисунку 2.10 помилки класифікатора діляться на два типи, помилки першого та другого роду.

Рисунок 2.10 – Матриця помилок

Продемонструвати їх роботу добре можна на прикладі з хворим та тестом який передбачує хворобу.

Якщо людина справді хвора, а тест показує що ні то це помилка другого роду (False Negatives).

Якщо ж людина не хвора, а тест показує що хвора то це помилка першого роду (False Positives).

Також дослідницькі дані показують що точність прогнозів дуже часто залежить від збалансованості прогнозованих класів у вибірці.

Тому існують ще дві метрики які покликані показані продемонструвати якість роботи класифікатора в умовах незбалансованості. Це метрики повноти (recall) та чутливості (precision).[28]

Обраховуються вони за формулами:

$$Precision = \frac{TP}{TP + FP};$$

$$Recall = \frac{TP}{TP + FN},$$

де TP - кількість правильно прогнозованих позитивних варіантів;

FP та FN - це помилки першого та другого роду.

Є також метрика яка їх об'єднує, яка називається F_1 і обчислюється за формулою:

$$F_1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

Метрика яка найбільш повно розкриває якість роботи класифікатора це площа (Area Under Curve) під кривою помилок (Receiver Operating Characteristic curve), або скорочено ROC AUC.[29-31]

Дана крива являє собою лінію від (0,0) до (1,1) в координатах True Positive Rate (TPR) та False Positive Rate (FPR), які обраховуються за формулами:

$$TPR = TP/(TP + FN)$$

$$FPR = FP/(FP + FN)$$

TPR вже нам відома, це повнота (recall), а FPR показує яку частку об'єктів з негативного класу алгоритм спрогнозував невірно.

В ідеальному випадку, коли класифікатор не робить помилок (FPR = 0, TPR = 1) отримаємо площу під кривою, що дорівнює одиниці. В іншому випадку, коли класифікатор випадково видає ймовірності класів (підкидання монети), AUC-ROC буде наближатись до 0.5, так як класифікатор видаватиме однакову кількість TP і FP.

Кожна точка на графіку відповідає вибору деякого порогу. Площа під кривою в даному випадку показує якість алгоритму (більше - краще), крім цього, важливою є крутизна самої кривої - ми хочемо максимізувати TPR, мінімізуючи FPR, крива помилок в ідеалі повинна прагнути до точки (0,1).

2.6 Висновок до розділу 2

У цьому розділі було проведено огляд існуючих методів, алгоритмів та моделей машинного навчання, які були використані для оцінювання платоспроможності у дослідженні, а саме логістична регресія, метод опорних векторів та випадковий ліс.

Показано один з принципів попередньої обробки даних, а саме балансування вибірки, наведені його види та описані їх переваги й недоліки.

Показано загальні метрики для оцінки якості моделей. Розглянуті та описані такі метрики як точність (accuracy), повнота (recall), чутливість (precision) та площа під кривою помилок (AUC ROC).

РОЗДІЛ 3. МОДЕЛЮВАННЯ ПЛАТОСПРОМОЖНОСТІ ПОЗИЧАЛЬНИКА

Основним завданням цієї роботи було проведення моделювання оцінювання платоспроможності методами машинного навчання, результатами якого можуть скористатись різні фінансові установи.

Це моделювання проводилось у декілька етапів:

1. підбір моделей для моделювання;
2. підготовка даних до моделювання;
3. виокремлення корисних ознак;
4. реалізація моделей;
5. запуск моделей на підготовлених даних;
6. збір результатів.

Практична частина дослідження була проведена за допомогою мови програмування Python 3.6 та бібліотек Scikit-learn та Pandas,оскільки цей інструментарій найкраще підходить для поставленої задачі, також для проектів машинного навчання та штучного інтелекту. Python надає доступ до бібліотек та фреймворків для машинного навчання (Machine Learning). Проекти на Python легкі у реалізації, та можуть бути запущені на будь-якій популярній сьогодні операційній системі.

Також потрібно зазначити що дослідження проводилось на персональному комп’ютері з процесором AMD Ryzen 5 3600 6-Core Processor 3.59 ГГц та 8 ГБ ОЗУ.

sklearn — це бібліотека Machine Learning для мови програмування Python, яка дає можливості для створення та тренування різних алгоритмів та моделей інтелектуального аналізу даних.

pandas — фреймворк, створений для мови програмування Python для роботи з даними та їх аналізу. Він дозволяє працювати з файлами як з структурами. За допомогою pandas можна легко та зручно обробляти великі

файли. Також перевагою серед інших інструментів є багато можливостей для аналізу файлових баз даних та часових рядів.

3.1 Опис вхідних даних

Вхідні дані - це анонімізована база даних користувачів Amazon Web Services. Зібрана вона була у 2017 році з понад 1 мільйона кредитів. Дані є анонімізовані отже жодних авторських прав не було порушено.

Інформація зберігається у двох файлах:

1. manualTransactions_funded_all.rpt - файл з даними про транзакції по кредитам;
2. all.rpt - файл з інформацією по самих кредитах, тобто ідентифікатор, тип кредиту, дата кредиту та багато іншої технічної інформації.

На рисунку 3.1 показано структуру файлу з транзакціями.

```
RangeIndex: 148215662 entries, 0 to 148215661
Data columns (total 4 columns):
advanceID         object
txnDescription    object
txnAmount         object
txnDate           object
dtypes: object(4)
memory usage: 4.4+ GB
```

Рисунок 3.1 – структура файлу з транзакціями

Він являє собою базу даних по кредитним транзакціям,яка містить інформацію по тому як були погашені ці кредити, дати, технічна інформація по платежам, суми транзакцій. Будемо використовувати цей файл як джерело ознак для тренувального датасету. Він також є цікавим з практичної точки

зору. База даних транзакцій є досить місткою, займає майже 17 ГБ дискового простору.

Як показано на рисунках 3.1 - 3.4 сутність транзакції складається з 4 елементів:

1. ідентифікатора;
2. суми;
3. опису транзакції;
4. дати транзакції.

	advanceID \
38465962	CBA6CFB1-4A05-E611-9454-02CC73CEC268
52777387	2C72DBE3-8FBA-E611-945B-02E98382A30D
144216047	75558693-5D9D-E611-945A-02CC73CEC268
130710971	B5D97FE1-FCE1-E711-945D-02E98382A30D
29752819	E2E3DA57-EA3B-E511-9451-02CC73CEC268

Рисунок 3.2 – Структура транзакції - Ідентифікатор

	txnDescription \
38465962	BASS HILL CUT PRICE 0001 BASS HILL Cash Out $28.50 Purchase $0.60
52777387	Returned Item CreditEffective Date: 29/09/2016
144216047	EFTPOS STEPPING STONE HAPPY VALLEY SA AU
130710971	Adjust Purchase ADJUSTMENT TO ACCOUNT Card xx5042 AUD 50.06 Value Date: 16/09/2017
29752819	ANZ M-BANKING PAYMENT TRANSFER xx4884 TO ANDREW FISHER

Рисунок 3.3 – Структура транзакції - Призначення платежу

	txnAmount	txnDate
38465962	-29.1	2016-02-09 18:00:00.000
52777387	213.33	2016-09-30 17:00:00.000
144216047	-273.41	2016-10-12 17:00:00.000
130710971	50.06	2017-11-22 18:00:00.000
29752819	-50	2015-05-13 17:00:00.000

Рисунок 3.4 – Структура транзакції - сума та дата транзакції

Другий файл містить інформацію по самим кредитам, а саме:

1. ідентифікатор, які співпадають з ідентифікаторами з першого файлу;
2. кількість кредитів, які вже були у даного позичальника;
3. тип кредиту (довгостроковий короткостроковий);
4. дата відкриття позички;
5. інформація про затримки у виплаті;
6. дані про те чи був списаний кредит.

На рисунках 3.5-3.6 показана структура цього файлу:

	advanceID	LoanNum	ProductName	\
913379	B1E41B64-01D0-E711-945D-02E98382A30D	2.0	SACC4%	
832502	3124E820-3244-E711-945B-02E98382A30D	13.0	SACC4%	
73074	4875EE08-9A63-E311-A835-005056A3762D	1.0	Loan*	
31190	FFFCBAD6-8229-E311-A835-005056A3762D	7.0	Loan*	
883845	10A5E7FA-BC9E-E711-945C-02E98382A30D	2.0	SACC4%	

Рисунок 3.5 – Структура кредиту - Ідентифікатор, кількість кредитів та тип.

	AGDateCreated	Fails35NoPayIn90	LoanWriteOff
913379	2017-11-23 13:50:06.140	0.0	0.0
832502	2017-05-29 15:46:17.133	0.0	0.0
73074	2013-12-13 11:58:22.090	0.0	0.0
31190	2013-09-30 13:46:16.270	0.0	0.0
883845	2017-09-21 21:07:04.937	0.0	0.0

Рисунок 3.6 – Структура кредиту - дата створення, просрочка по виплатам та списання.

Як можна помітити кредити містять 6 змінних. Цільовою змінною тут буде виступати FailPaysIn90, оскільки вона містить інформацію про те чи був закритий даний кредит.

Як показано на рисунку 3.7, вибірка є незбалансованою. Кількість кредитів, що були нормально закриті значно перевищує кількість проблемних. Кількість невиплачених кредитів складає лише 24501, коли кількість вчасно погашених складає 432637.

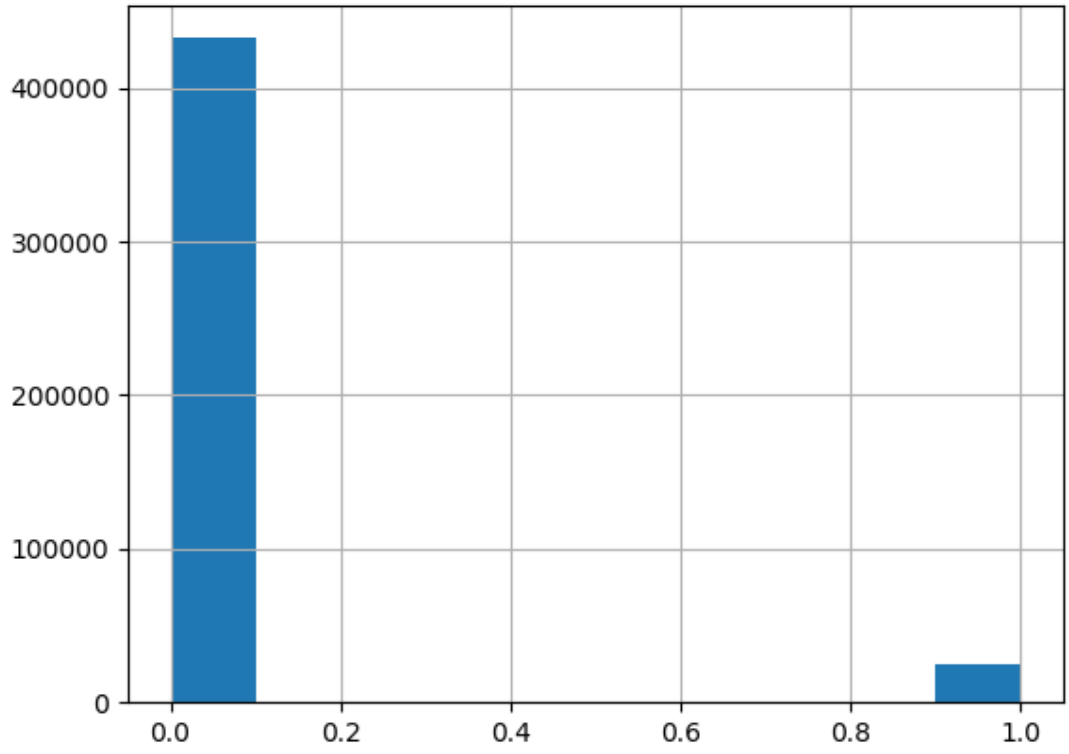

Рисунок 3.7 – Розподіл класів у тренувальній вибірці

Як бачимо кількість “гарних” випадків приблизно у 20 раз більша ніж кількість поганих.

3.2 Попередня обробка даних

Перед початком моделювання дані потрібно обробити. Слід зазначити що буде проводитись бінарна класифікація.

Оскільки система на якій проводилось моделювання була доволі таки обмежена в ресурсах було вирішено розділити файл з транзакціями на окремі файли в кожному з яких міститься по одному мільйону транзакцій. Потім вже працювати з окремими частинами цих транзакцій.

Наступним етапом, стало виокремлення корисних для моделей ознак. Такими ознаками вирішено було взяти: статистичні дані по окремим кредитам:

1. транзакція;
2. мінімальна максимальна транзакція;
3. сума всіх транзакцій;
4. їх середнє арифметичне.

Оскільки дані не ідеально структуровані, та містять велику кількість записів, серед них було виявлено багато проблемних рядків. Такі рядки було вирішено видалити.Це продемонстровано на рисунках 3.8-3.10

```
Number of errors: 12326
```

Рисунок 3.8 – Кількість помилок при читанні файлу

```
df = df.fillna('', inplace=False)
df = df[df['txn_max'] != '']
df = df[df['txn_mean'] != '']
df = df[df['txn_min'] != '']
df = df[df['txn_sum'] != '']
```

Рисунок 3.9 - Видалення порожніх значень

```
Number of errors: 0
```

Рисунок 3.10 – Кількість помилок після процедури видалення

Далі була поставлена задача об’єднати вибірки з корисними ознаками та вибірки з результатами кредитних виплат.

Останнім етапом попередньої обробки даних стало створення вже кінцевої вибірки даних. На рисунку 3.11 показано властивості вибірки після обробки даних.

```
Int64Index: 457138 entries, 0 to 458758
Data columns (total 6 columns):
advanceID           457138 non-null object
txn_max             457138 non-null float64
txn_mean            457138 non-null float64
txn_min             457138 non-null float64
txn_sum             457138 non-null float64
Fails35NoPayIn90    457138 non-null int32
dtypes: float64(4), int32(1), object(1)
memory usage: 22.7+ MB
```

Рисунок 3.11 – Опис тренувальної вибірки

Як бачимо змінні тепер строго визначені як числові, що необхідно для коректної роботи більшості класифікаторів. Також можна помітити що кількість кредитів зменшилась майже в двічі, це пов’язано з тим що цільовий файл по кредитам містив багато зайвих кредитів, яких не було у файлі з транзакціями.

3.3 Використання логістичної регресії

Перед навчанням моделі розіб’ємо нашу вибірку на тренувальну і тестувальну. Оскільки вибірка дуже значна по об’єму то спробуємо розбити у співвідношенні 25% на тест, та 75% на тренувальну.

Як показано на рисунку 3.12 в результаті отримали значення ROC AUC 0.51 та точність прогнозу:0.9462487904305038.

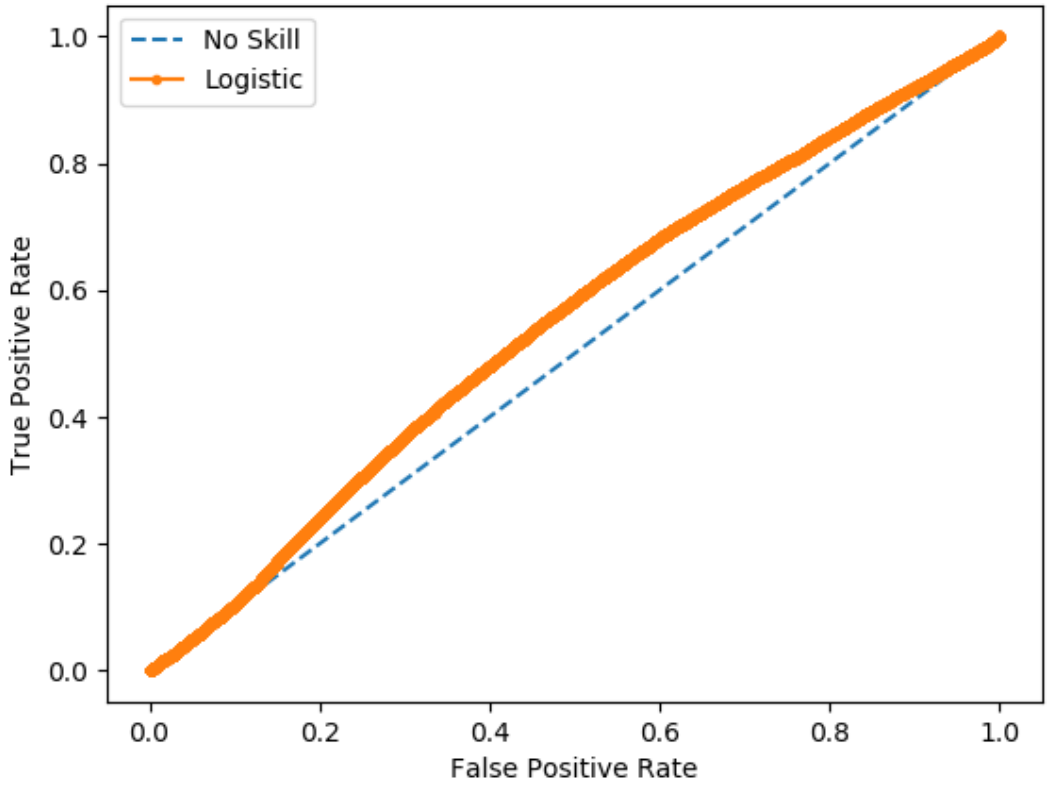

Рисунок 3.12 - Результат роботи моделі

Як бачимо точність прогнозу висока, але крива ROC показала що результат не набагато краще підкидання монетки. Це зв'язано з тим що вибірка є незбалансованою і дана модель передбачила все як 0 і отримала високий результат точності. І ця гіпотеза підтверджується якщо переглянути прогноз детальніше на рисунку 3.13.

```
0    388559
1         9
```

Рисунок 3.13 - Результат прогнозу

Тобто майже всі рішення що приймала модель - це відмовити людям в кредитах.

Спробуємо провести балансування вибірки та повторити експеримент. Було проведено балансування методом додавання прикладів. Параметри наведені на рисунку 3.14.

```
df_minority_upsampled = resample(df_minority, replace=True,
                                 n_samples=430000, random_state=123)
```

Рисунок 3.14 - Параметри балансування

Після балансування кількість “поганих” та “хороших” кредитів стала майже однаковою. В цьому можна переконатись проаналізувавши рисунок 3.15. Це зроблено для покращення якості прогнозу моделей.

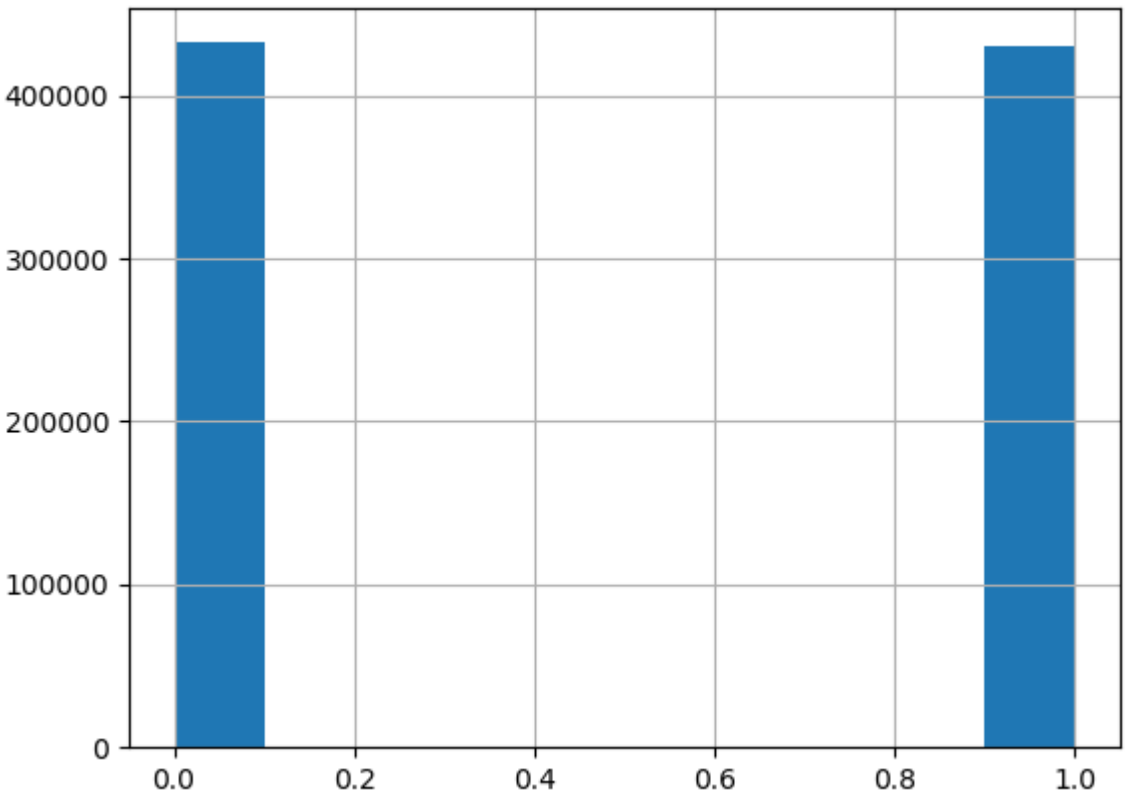

Рисунок 3.15 - Вибірка після балансування

Також можна побачити що після балансування вибірки значно погіршилась точність прогнозу і становить приблизно 0.53, проте тепер

модель передбачає не тільки 0. Також покращилась і метрика ROC і становить тепер 0.55. Це показано на рисунку 3.16.

З цього прикладу видно що для класифікаторів проблема балансування класів є дуже значною. І навіть такий простий прийом як додавання вже існуючих прикладів до вибірки значно покращує результати даної моделі.

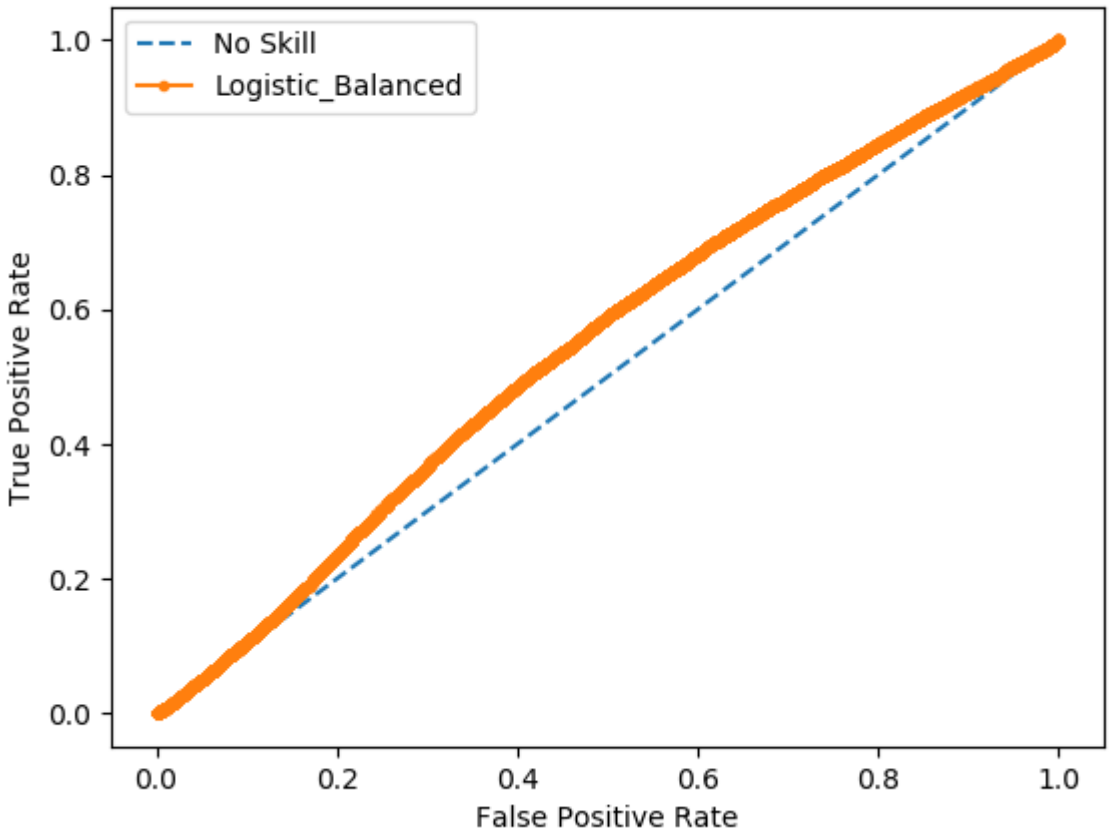

Рисунок 3.16 - ROC AUC після балансування

Отже, логістична регресія без балансування вибірки показала невтішні результати. Але при використанні балансування ця модель стала відображати реальний стан речей і дещо покращила значення метрик.

3.4 Моделювання методом опорних векторів

Необхідною складовою коректної роботи класифікатора є нормалізація даних. Для нормалізації було використано MinMaxScaler, який всі значення у вибірці ділив на максимальне значення, так щоб вся вибірка перебувала у діапазоні (0,1). На рисунка 3.17-3.18 показана ініціалізація методу опорних векторів.

```
svc = OneVsRestClassifier(SVC(kernel='linear',
                              class_weight='balanced', probability=True,tol=1e-3), n_jobs=-1)
```

Рисунок 3.17 - Параметри методу опорних векторів

```
X_train, X_test, y_train, y_test = train_test_split(X, y,
                                                    test_size=0.15,
                                                    random_state=42)
```

Рисунок 3.18 - Параметри розбиття вибірки для навчання на SVM

Отже, функцією ядра було обрано лінійну, оскільки вибірка вже збалансована то ваги для класів встановлено як збалансовані, увімкнено режим підрахунку ймовірностей, точність встановлена як 0.001. І для пришвидшення обрахунків було застосовано параметр n_jobs як -1 для використання всіх доступних ресурсів процесора.

Результати першого запуску були не дуже вражаючими: точність прогнозу 0.497, roc auc 0.495.

В таблиці 3.1 наведено результати роботи алгоритму залежно від, функції ядра, точності навчання та параметру багатопоточності. Помітно, що значний приріст у часі супроводжується досить незначними покращеннями у результаті.

Таблиця 3.1 – Опис результатів роботи SVM залежно від значень параметрів.

Функція ядра	Точність навчання	Параметр багатопоточності	Час роботи в секундах	Точність прогнозу	ROC
Лінійна	0.001	0	3647	0.5268	0.5113
Лінійна	0.001	1	1956	0.5268	0.5113
Лінійна	0.01	1	477	0.4973	0.4956
Поліноміальна	0.001	0	4669	0.5229	0.5184
Поліноміальна	0.001	1	2194	0.5229	0.5184
Поліноміальна	0.01	1	590	0.4973	0.5017

Також незадовільними залишаються і якісні показники прогнозу даного алгоритму. Як показано на рисунку 3.19 модель показала досить посередні результати, в деяких варіаціях навіть гірші, ніж при випадковому вгадуванні.

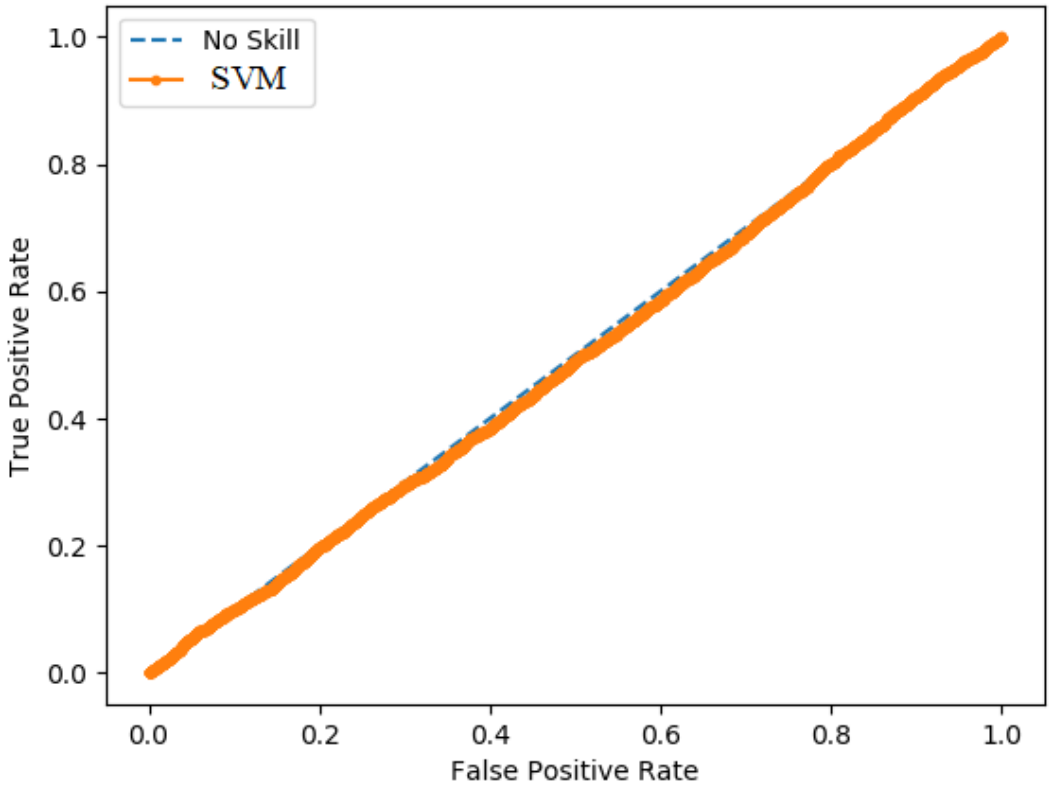

Рисунок 3.19 - Результат роботи SVM

Окрім поганої пристосованості до сучасних задач, у цього методу є ще один великий недолік у вигляді значних потреб у оптимізації.

Намагаючись зрозуміти причини невдачі цього алгоритму на даній задачі було зроблено певні висновки.

Дані дуже рідко подаються у вигляді векторів характеристик, як того потребує метод опорних векторів. Для коректної роботи SVM вимагає певного методу вилучення особливостей, щоб отримувати гарні результати. При цьому якщо SVM отримує результати іншого методу, тобто застосовується метод ансамблювання, і працює як фінальний класифікатор, то його результати значно кращі. Але при роботі окремо, цей метод не може показати значних результатів.[23]

3.5 Моделювання методом випадковий ліс (Random forest)

Наступним методом використаним для моделювання є Random Forest. Важливим аспектом застосування цього алгоритму є те, що на великих вибірках він дуже схильний до перенавчання. На рисунку 3.20 показано параметри ініціалізації методу Random Forest.

```
clf_2 = RandomForestClassifier(n_estimators = 100,
                               criterion="gini").fit(X_train, y_train)
```

Рисунок 3.20 – Параметри Random Forest.

Серед параметрів було обрана кількість “дерев” - 100 та критерій gini як статистичну метрику алгоритму. Ці параметри дозволяють значно покращити результати.

На рисунках 3.21 та 3.22 показані результати першої спроби Random Forest. Моделювання проводилось з стандартними параметрами розбиття вибірки на тренувальну і тестову як 75% та 25% виникла явна проблема перенавчання. Точність прогнозу досягла 99.7%. Це означає що при виході за межі даної вибірки модель не буде здатна адекватно працювати.

```
0.997264212185848
1     64769
0     64627
```

Рисунок 3.21 - Результати прогнозу Random Forest - точність та розподіл по класам.

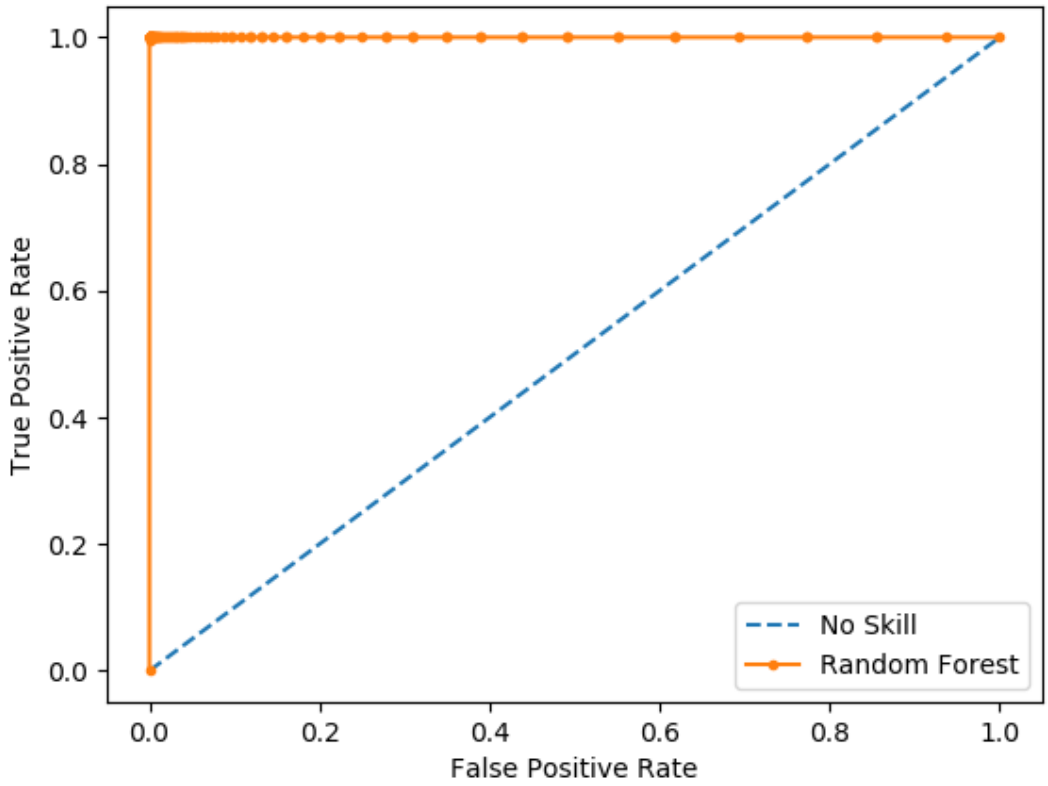

Рисунок 3.22 - Roc Auc при перенавчанні

Для запобігання перенавчання було розбито вибірку як показано на рисунку 3.23 так щоб тренувальна частина склала лише 15% від її.

```
X_train, X_test, y_train, y_test = train_test_split(X, y,
                                                    test_size=0.85,
                                                    random_state=42)
```

Рисунок 3.23 - Параметри для розбиття вибірки

Після спроби запобігти перенавчанню результати краще ніж у попередніх моделей. На рисунку 3.24 можна помітити, що точність прогнозу досягла 90% та метрика roc auc більше 0.96.

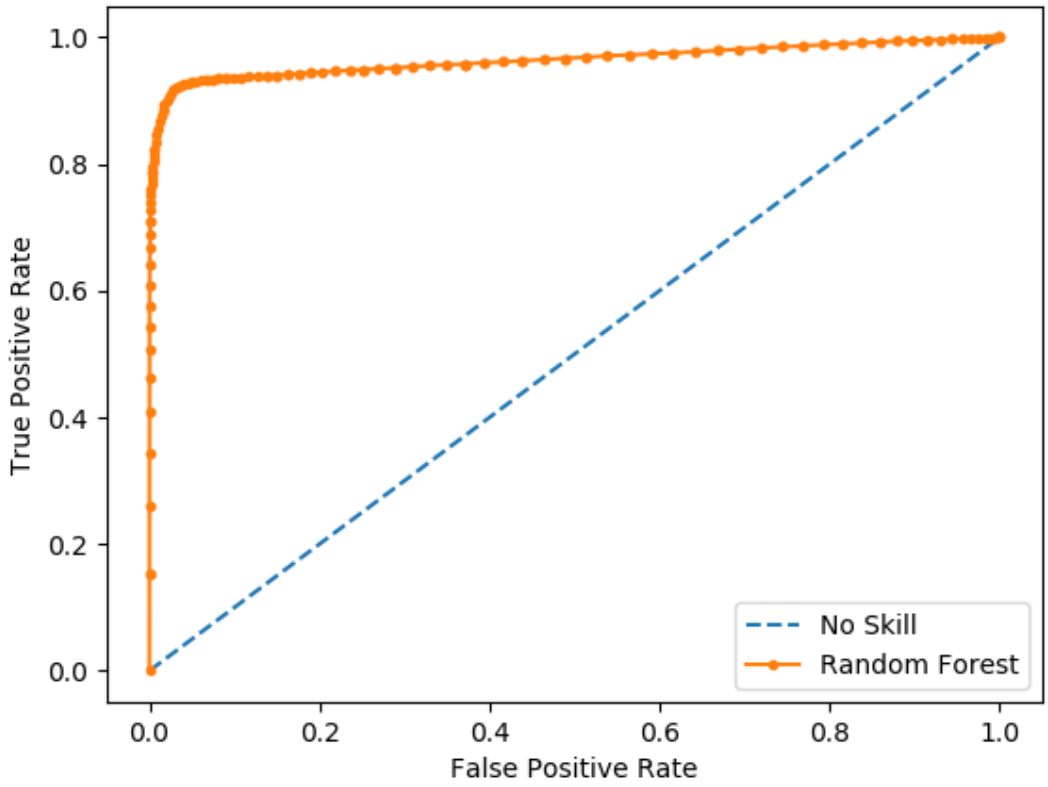

Рисунок 3.24 - Графік roc auc для Random Forest при запобіганні перенавчання

3.6 Порівняння результатів

Заключним етапом дослідження є порівняння результатів різних моделей, та розробка висновків.

В таблиці 3.2-3.4 наведено результати роботи всіх класифікаторів. Потрібно зазначити що до таблиці додано найкращі результати роботи класифікаторів за певних умов(збалансованість даних, запобігання перенавчання).

Таблиця 3.2 – Порівняння результатів роботи класифікаторів на незбалансованих даних та без запобігання перенавчання.

Назва	Точність(accuracy)	ROC	Час роботи, с.
LogReg	0.94	0.54	35
SVM	0.93	0.51	2717
Random Forest	0.997	0.999	28

Таблиця 3.3 – Порівняння результатів роботи класифікаторів на збалансованих даних та без запобігання перенавчання.

Назва	Точність(accuracy)	ROC	Час роботи, с.
LogReg	0.55	0.53	55
SVM	0.52	0.51	4723
Random Forest	0.997	0.999	53

Таблиця 3.4 – Порівняння результатів роботи класифікаторів на збалансованих даних та при запобігання перенавчання.

Назва	Точність(accuracy)	ROC	Час роботи, с.
LogReg	0.55	0.53	55
SVM	0.52	0.51	4723
Random Forest	0.90	0.96	47

Потрібно зазначити що запобігання перенавчання проводилось тільки для методу випадкових лісів, та в таблиці наведено як порівняння результатів з іншими методами.

Як можна помітити найкращим результат показує модель Random Forest, хоч і є проблеми з перенавчанням, але якщо спробувати їх подолати зменшивши відсоток даних тренувальної вибірки відносно всіх даних, то отримуємо покращений результат.

Отже отримано результат який показує значні переваги методу Random Forest перед іншими моделями, як по часовим затратам так і по точності прогнозу. Також треба звернути увагу що час роботи моделей здебільшого складається з часу їх навчання та при застосування в реальних умовах, тобто коли моделі вже навчені час роботи буде значно меншим.

3.7 Висновки до розділу 3

В цьому розділі було проведено практичне моделювання.

Було проаналізовано та оброблено дані по кредитам за 2017 рік. Кредитів виявилось близько мільйона та ще 149 мільйонів транзакцій по ним. В результаті обробки даних транзакцій та кредитів їх кількість зменшилась до півмільйона кредитів. Потім було навчено 3 моделі для прогнозування дефолту позичальників: логістична регресія, метод опорних векторів та випадковий ліс. Результати моделювання оцінювались метриками точності (accuracy) та метриками roc (receiver operating characteristic). Були продемонстровані графіки для результатів кожної моделі у вигляді кривої помилок.

В результаті практичної частини дослідження логістична регресія показала посередній результат та виявилась непридатною для використання в даній задачі без балансування вибірки(точність 0.55). Наступним алгоритмом

став метод опорних векторів, який виявився взагалі непридатний до використання в сучасних умовах. По перше дуже великі часові затрати на навчання при досить незначній точності моделі.(точність 0.495). Найкращий результат показав метод Random Forest. При моделюванні цим методом виникла проблема з проблемою значного перенавчання, але при її подоланні цей метод показав дуже серйозні результати (точність 0.90, roc auc 0.96). Високе значення roc auc означає що кількість помилок першого та другого роду значно зменшилась в порівнянні з попередніми моделями.

РОЗДІЛ 4 РОЗРОБКА СТАРТАП ПРОЕКТУ

На сьогодні великої популярності набуває такий вид підприємництва як стартап. Стартап-проект — є комерційним проектом, який знаходиться в стані розробки, або нещодавно вийшов на ринок. Характерною особливістю стартапу, що відрізняє його від малого бізнесу, є оригінальність та інновації, він не може бути копією вже реалізованих ідей. При цьому проект не обов’язково повинен бути масштабного характеру, головне, щоби він був креативним, а його завдання — спрощувати людям будь-які дії в їх повсякденному житті.

Наразі, з появою Інтернету та сучасних технологій, стало простіше заходити на ринок, знаходити інвесторів та споживачів. З’явилося набагато більше можливостей для розвитку свого проекту за кордоном, ніж раніше.

Проте розробка стартапу є досить ризикованим завданням. Не всім вдається довести свій стартап-проект до ринкового впровадження. За статистикою успіху досягає лише 10–20 % від усіх стартап-проектів.

Запуск стартапу передбачає цілий ряд обов’язкових дій, у межах яких визначають ринкові перспективи стартапу, графік розробки, принципи організації виробництва, заходи з залучення інвесторів та аналіз ризиків.

4.1 Опис ідеї проекту

У таблиці 4.1 описані основні ідеї проекту, можливі напрямки застосування та основні вигоди, що може отримати користувач товару. У таблиці 4.2 визначені сильні, слабкі та нейтральні сторони проекту.

Таблиця 4.1— Опис ідеї стартапу.

Зміст ідеї	Напрямки застосування	Вигоди для користувача
Програмний продукт для банківських установ для швидкої та якісної оцінки платоспроможності потенційних клієнтів	Банки та інші фінансово-кредитні установи	Дозволяє проводити швидку обробку даних для подальшої побудови прогнозуючих моделей.

Таблиця 4.2 — Порівняльна характеристика ідеї

Техніко-економічні характеристики ідеї	Мій проект	IBM SPSS Modeler	SAS Enterprise Miner
Ціна	Низька	Висока	Висока
Функціонал	Вузький	Широкий	Широкий

Отже з таблиці видно що сильною стороною є ціна, але функціонал не дозволяє замінити більш відомі конкурентні програмні продукти.

4.2 Технологічний аудит проекту

Наступним етапом стартап-аналізу є технологічний аудит. Його метою є порівняння продукту що розробляється з іншими доступними на ринку продуктами.

В межах даного підрозділу проведено аудит технології, за допомогою якої реалізовано ідею проекту.

Визначення технологічної здійсненності ідеї проекту передбачає аналіз таких складових (таблиця 4.3):

- за якою технологією буде виготовлено товар згідно ідеї проекту?
- чи існують такі технології, чи їх потрібно розробити/доробити?
- чи доступні такі технології авторам проекту?

Таблиця 4.3 — Визначення сильних, слабких та нейтральних характеристик ідеї проекту

Техніко-економічні характеристики ідеї	*(потенційні) товари/концепції конкурентів*		
	Мій проект	*PhLAb*	*Lokad*
Точність прогнозування	Кілька моделей з різними результатами	Власний алгоритм	Алгоритм *Lokad*
Складання ймовірності погашення кредиту	Присутня, але точність не найкраща.	Дана функція повністю відсутня	Дана функція повністю відсутня
Ризики невірного прогнозу	Існують, через велику кількість факторів.	Невідомо	Відсутні через відсутність пронозу
Доступність, зручність	Десктопний застосунок	Власний інтерфейс	Власний інтерфейс

4.2.1 Технологічна здійсненність ідеї проекту

В таблиці 4.4 показана технологічна здійсненність ідеї проекту. Ця таблиця показує чому серед аналогічних мов програмування, для дослідження була обрана саме мова програмування Python.

Таблиця 4.4 — Технологічна здійсненність ідеї проекту

Ідея проекту	*Технології її реалізації*	*Наявність технологій*	*Доступність технологій*
Створення системи оцінки ймовірності повернення кредиту	Використання мови C++	Відсутні	Недоступні
	Використання мови програмування C#	Відсутні	Недоступні
	Використання мови програмування Python	Наявна	Доступні
Обрана технологія реалізації ідеї проекту: мова програмування python			

4.2.2 Попередня характеристика потенційного ринку

Значної уваги треба приділити характеристиці ринку (Таблиця 4.5), оскільки подальший ефект від впровадження залежить від зацікавленості основних клієнтів та зацікавлених сторін.

Таблиця 4.5 — Попередня характеристика потенційного ринку автоматизованих прогнозів по поверненню кредитів.

Показники стану ринку (найменування)	*Характеристика*
Кількість головних гравців, од	50
Загальний обсяг продаж, грн/ум.од	100000
Динаміка ринку (якісна оцінка)	Зростає
Наявність обмежень для входу (вказати характер обмежень)	Немає
Специфічні вимоги до стандартизації та сертифікації	Немає
Середня норма рентабельності в галузі (або по ринку),%	17 %

4.2.3 Характеристика потенційних клієнтів

Окремо треба розглядати особливості клієнтів (Таблиця 4.6) з метою визначення цільової аудиторії (цільові сегменти ринку) та відмінності у поведінці різних потенційних цільових груп клієнтів.

Таблиця 4.6 — Характеристика потенційних клієнтів стартап-проекту

Потреба, що формує ринок	Цільова аудиторія (цільові сегменти ринку)	Відмінності в поведінці різних потенційних цільових груп клієнтів	Вимоги споживачів до товару
Статистичні параметри роботи моделей, місячний прогноз	Аналітики, аналітичні відділи	Низька ціна, велика кількість статистичних даних.	Простота використання.
Створення якісного прогнозу	Малі фінансові установи	Цікавить простота у використанні, низька ціна клієнська підтримка	Низька ціна, репутація
Створення точного прогнозу, та постійна довгострокова підтримка	Великі компанії, Банки	Цікавить передусім точність довгострокових прогнозів, клієнська підтримка	Висока якість, бренд, ім’я на ринку, успішний досвід

4.2.4 Фактори загроз

Подальше впровадження проекту тісно пов’язано з факторами загроз, оскільки саме вони дозволяють виявити доцільність реалізації проекту (Таблиця 4.7).

Після визначення потенційних груп клієнтів проведено аналіз ринкового середовища: складено таблиці факторів, що сприяють ринковому впровадженню проекту, та факторів, що йому перешкоджають. Фактори в таблиці подані в порядку зменшення значущості.

Таблиця 4.7 — Фактори загроз

Фактор	*Зміст загрози*	*Можлива реакція компанії*
Конкуренція	Цього року очікується вихід на ринок крупної іноземної компанії-конкурента	Пришвидшити вихід програмного продукту
Збут	Ускладнення збуту через цінову політику конкурентів	Розміщення додаткових рекламних банерів в Інтернеті

Також треба визначити фактори можливостей (Таблиця 4.8) та провести аналіз конкуренції (Таблиця 4.9, Таблиця 4.10).

Таблиця 4.8 — Фактори можливостей

Фактор	*Зміст можливості*	*Можлива реакція компанії*
Гнучкі ціни	Зменшення ціни товару задля збільшення попиту	Введення власних гнучких цін
Диференціація витрат	Зменшення витрат за рахунок їх перерозподілу	Зменшення витрат на додаткові, непрофільні задачі.

Таблиця 4.9 — Ступеневий аналіз конкуренції на ринку

Особливості конкурентного середовища	У чо*му проявляється дана характеристика*	*Вплив на діяльність підприємства (можливі дії компанії, щоби бути конкурентоспроможною)*
1. Вказати тип конкуренції— Досконала конкуренція	Багато систем/команд аналітиків	Розробити впізнаваний продукт, якість, що вирізнятиме нас від конкурентів
2. За рівнем конкурентної боротьби: міжнародний	На ринку присутні системи, розроблені за кордоном.	Розширення аудиторії, розширення списку мов, які підтримуються системою
3. За галузевою ознакою— міжгалузева	Робота із НБУ в різних галузях.	Розширення списку спеціалізованих моделей, де система може бути застосована
4. Конкуренція за видами товарів: товарно-родова	Конкуренція між прогнозами інформаційних систем та команд аналітиків.	Збільшення функціоналу системи
5.За характером конкурентних переваг: Нецінова	Різні способи прогнозування дають різну точність	Розробка кращих(точніших) алгоритмів
6. За інтенсивністю: Марочна	Впізнаваний бренд надає великих переваг	Велику увагу приділити розвитку бренду

4.2.5 Обґрунтування факторів конкурентоспроможності

Проведений аналіз дозволяє виявити фактори конкурентоспроможності (Таблиця 4.10).

Таблиця 4.10 — Обґрунтування факторів конкурентоспроможності

Фактор конкурентоспроможності	*Обґрунтування (наведення чинників, що роблять фактор для порівняння конкурентних проектів значущим)*
Багатофункціональність	Жоден конкурент не є настільки багатофункціональним, не здатен на прогноз та порівняння результату моделей.
Якість	Висока якість прогнозу, велика кількість допоміжних статистичних даних
Обслуговування	Робота з клієнтами — передусім великим та середнім бізнесом

4.2.6 SWOT- аналіз стартап-проекту

Для проведення аналізу можуть залучатись різноманітні методи, зокрема SWOT – аналіз (Таблиця 4.11).

Таблиця 4.11 — SWOT- аналіз стартап-проекту

Сильні сторони: Висока якість прогнозу, якість, багатофункціональність	Слабкі сторони: відсутній, інтерфейс користувача, немає налагодженої клієнтської бази
Можливості: Попит, зміна рівня доходів населення, вдосконалення системи	Загрози: конкуренція, збут

Дані з альтернативи ринкового впровадження наведено у таблиці 4.12.

Таблиця 4.12 — Альтернативи ринкового впровадження стартап-проекту

Альтернатива (орієнтовний комплекс заходів) ринкової поведінки	*Ймовірність отримання ресурсів*	*Строки реалізації*
Швидкий вихід на ринок із «сирим» продуктом, можливі проблеми із точністю прогнозу та універсальністю	30 %	3 місяці
Поступовий вихід із готовим, відлагодженим продуктом. Висока якість та конкурентоспроможна ціна.	70 %	6 місяців

4.3 Розроблення ринкової стратегії проекту

4.3.1 Вибір цільових груп потенційних споживачів

Ключовими моментами розроблення ринкової стратегії проекту є групи потенційних споживачів (Таблиця 4.13), стратегія розвитку (Таблиця 4.14),

Таблиця 4.13 — Вибір цільових груп потенційних споживачів

п/п	*Опис профілю цільової групи потенційних клієнтів*	*Готовність споживачів сприйняти продукт*	*Орієнтовний попит у межах цільової групи (сегменту)*	*Інтенсивність конкуренції в сегменті*	*Простота входу в сегмент*

Продовження таблиці 4.13

1	Окремі аналітики та аналітичні відділи	Низька готовність	25 %	Висока	Середня
2	Малі та середні аутсорсингові контакт центри	Висока	40 %	Середня	Висока
3	Великі компанії із власними кол-центрами	Висока	20 %	Низька	Середня

Таблиця 4.14 — Визначення базової стратегії розвитку

Обрана альтернатива розвитку проекту	Стратегія охоплення ринку	Ключові конкурентоспроможні позиції відповідно до обраної альтернативи	Базова стратегія розвитку*
2 та 3	Стратегія диференційованого маркетингу	Висока універсальність, багатогалузевість, висока якість, ціна.	Стратегія лідерства по витратах

4.3.2 Визначення базової стратегії конкурентної поведінки

В таблиці 4.15 наведена інформація про базові стратегії конкурентної поведінки.

Таблиця 4.15 — Визначення базової стратегії конкурентної поведінки

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	Так	Ні	Стратегія виклику лідера

4.3.3 Визначення стратегії позиціонування

Важливим етапом стартап-аналізу проекту є визначення стратегії позиціонування. Вона дозволяє зрозуміти яким чином потрібно позиціонувати товар на ринку. Наведена стратегія у таблиці 4.16

Таблиця 4.16 — Визначення стратегії позиціонування

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)*
Якість, точність, простота у використанні	Стратегія лідерства по витратах	Якість прогнозу, універсальність, велика кількість статистичної інформації	По іміджу. Позиціювання на низькій ціні. Позиціювання за сферою застосування

4.4 Розроблення маркетингової програми

4.4.1 Ключові переваги

Особливу увагу треба приділити визначенню ключових переваг концепції потенційного товару в таблиці 4.17.

Таблиця 4.17 — Визначення ключових переваг концепції потенційного товару

<table>
<tr><td>Потреба</td><td>Вигода, яку пропонує товар</td><td colspan="3">Ключові переваги перед конкурентами (існуючі або такі, що потрібно створити</td></tr>
<tr><td>Статистичні параметри роботи поведінкової моделі, місячний прогноз</td><td>Розгорнута статистика, яка охоплює всі особливості роботи поведінкової моделі.</td><td colspan="3">Дана статистика має бути максимально розгорнутою та зрозумілою.</td></tr>
<tr><td>Створення якісного та точного прогнозу</td><td>Точний прогноз — основа функціонування будь-якого Банку.</td><td colspan="3">Розробка коротко- та довгострокових прогнозів, використання різних методів для покращення точності прогнозу</td></tr>
<tr><td rowspan="5">II. Товар у реальному виконанні</td><td>Властивості/характеристики</td><td>М/Нм</td><td>Вр/Тх /Тл/Е/Ор</td></tr>
<tr><td colspan="3">Програмний продукт — складна, комплексна система</td></tr>
<tr><td colspan="3">Якість — тестування сторонніми фірмами</td></tr>
<tr><td colspan="3">Пакування — зручний інсталятор</td></tr>
<tr><td colspan="3">Марка: назва організації-розробника + назва товару</td></tr>
<tr><td rowspan="2">III. Товар із підкріпленням</td><td colspan="3">До продажу: доставка, гнучкі умови оплати</td></tr>
<tr><td colspan="3">Після продажу: гарантія якості</td></tr>
<tr><td colspan="4">За рахунок чого потенційний товар буде захищено від копіювання: Система авторського права</td></tr>
</table>

4.4.2 Формування систем збуту

Для визначення межі цін (Таблиця 4.18) та системи збуту (Таблиця 4.19) можуть залучатись додаткові дані фінансового ринку.

Таблиця 4.18 — Визначення меж встановлення ціни

Рівень цін на товари-замінники	*Рівень цін на товари-аналоги*	*Рівень доходів цільової групи споживачів*	*Верхня та нижня межі встановлення ціни на товар/послугу*
10000	15000	3,00 млн	8000–9000

Таблиця 4.19 — Формування системи збуту

Специфіка закупівельної поведінки цільових клієнтів	*Функції збуту, які має виконувати постачальник товару*	*Глибина каналу збуту*	*Оптимальна система збуту*
Канал нульового рівня	Доставка товару	0	+

4.5 Висновки до розділу

В даному розділі розглянуто стартап-ідею проектую.

Основні цілі, що були досягнуті в цьому розділі:

1) з’ясовано, що є можливість ринкової комерціалізації проекту (наявні попит, динаміка ринку, рентабельність роботи на ринку);
2) з’ясовано, що є перспективи впровадження з огляду на потенційні групи клієнтів, бар’єри входження, стан конкуренції;
3) конкурентоспроможність проекту є високою;
4) подальша імплементація проекту є можливою.

З огляду на потенційні групи клієнтів та бар’єри, які стоять на шляху, можна сказати, що у даного проекту є досить непогані перспективи впровадження. При цьому рівень конкуренції на даний момент є не дуже високим, а конкурентоспроможність проекту є достатньою.

Для ринкової реалізації проекту, на даний момент, краще обрати варіант розробки продукту, при якому буде використовуватися мова програмування Python.

ВИСНОВКИ

Дана робота присвячена дослідженню в галузі машинного навчання для оцінювання платоспроможності. В першому розділі було проведено огляд історичних методів для оцінювання платоспроможності. Було описано такі методики, як PARSER, CAMPARI, 4FC, PARTS, MEMO RISK, 6C та розглянуто їх способи використання.

Дослідження полягало у тому щоб навчити моделі, методами машинного навчання, передбачати здатність позичальників до повернення взятих на себе коштів. Детальний опис методів та засобів для моделювання, обробки вхідних даних та способів оцінювання якості роботи моделей було проведено у другому розділі. До методів які описані у цьому дослідженні належать: логістична регресія, метод опорних векторів та метод випадковий ліс. Також для якісної їх роботи у другому розділі було детально описано методи балансування вибірки, оскільки при його відсутності більшість класифікаторів не досягне поставленого результату.

Третій розділ присвячений практичній частині цього дослідження. Там описана практична складова дослідження, показані основні елементи які використовувались при програмуванні на мові Python. Було проведено обробку вхідних даних, результатами якої стало формування вибірки для навчання і тестування моделей. За результатами тестування моделей виявилось, що жодна з розглянутих моделей не може коректно працювати у випадку незбалансованості даних по класам. Але після проведення балансування найкраще себе показав метод Random Forest, який продемонстрував точність прогнозу 90% та високий результат по метриці ROC 0.96. Наступним після цього методу йде логістична регресія з точністю на рівні 53% та ROC на рівні 0.55. І найгіршим методом для розв’язання даної задачі при схожих даних є метод опорних векторів. Який при значних витратах часу показав досить поганий результат з точністю 49,5% та roc 0.493 що означає

якщо робити навпаки до того що рекомендує цей алгоритм, то отримаємо кращий результат.

В четвертому розділі було проведено стартап-аналіз проекту за результатами якого було визначено, що проект є конкурентноздатним і можлива подальша інтеграція його на ринок.

Отже проведене дослідження відповідає всім поставленим завданням, та в майбутньому існують перспективи розвитку даного дослідження, наприклад аналіз більшої кількості моделей, побудова кроссплатформенного застосунку для впровадження його в роботу з клієнтами. Також можливе покращення дослідження шляхом збору більш актуальних даних.

ПЕРЕЛІК ПОСИЛАНЬ

1. Банківська енциклопедія. С.Г. Арбузов, та ін. Київ: Центр наукових досліджень Національного банку України «Знання», 2011. 503 с.
2. Кириченко О.А. Банківський менеджмент: Навч. Посібник. Київ «Знання-Прес». 2002. 438 с.
3. Бордюг В.В. Теоретичні основи оцінки кредитоспроможності позичальника банку. *Вісник Університету банківської справи Національного банку Україн*и. 2008. № 3. С. 112–115.
4. Steve Patterson. The 6 C's of Business Credit. *Finance and Economics*. 2013. №3. P. 18, URL: https://www.score.org/resource/6-cs-business-credit (дата звернення: 24.10.2020)
5. Смолева Т.М. Сучасні методи оцінки кредитоспроможності позичальників банками України. *Фінанси, облік, банки.* 2014. № 1. С. 241–245.
6. Кузнєцова Н. В. Порівняльний аналіз характеристик моделей оцінювання ризиків кредитування. *Наукові вісті НТУУ «КПІ»*,.2010, №1, С. 42–53.
7. Проф. С.В. Васильчак, д-р екон. наук; магістрант Л.Р. Демус. Оцінка кредитоспроможності позичальника як один із методів забезпечення економічної безпеки банку. *Науковий вісник НЛТУ України.* 2012. Вип. 22.1, с.154-161. URL: https://nv.nltu.edu.ua/Archive/2012/22_1/154_Was.pdf
8. Чорноморченко Н. В. Обґрунтування господарських рішень і оцінювання ризиків. Навч.-метод. посібник. Львів. Магнолія-2006. 2010. 260 с.
9. Шегда А. В. Ризики в підприємництві: оцінювання та управління : навч. посіб. Київ : Знання. 2008. 271 с.

10. Положення НБУ №64 про організацію системи управління ризиками в банках України та банківських групах: посібник. Київ: НБУ 2018. 105 с. URL: https://bank.gov.ua/document/download?docId=71600453 (дата звернення: 26.10.2020)
11. Іванілов О.С. Економіка підприємства: підручник. Київ: ЦУЛ, 2009. 728 с.
12. EISENBEIS, R. A.: Pitfalls in the Application of Discriminant Analysis in Business, Finance and Economics. *Journal of Finance*. 1977, vol. 32, P. 875-900.
13. Tobias Berg, Valentin Burg, Ana Gombović. On the Rise of FinTechs– Credit Scoring using Digital Footprints. *Federal Deposit Insurance Corporation, Center of Financial Research*. 2018, № 4, P. 27-28. URL:https://www.fdic.gov/bank/analytical/cfr/2018/wp2018/cfr-wp2018-04.pdf (дата звернення: 07.11.2020)
14. Mikella Hurley,Julius Adebayo. CREDIT SCORING IN THE ERA OF BIG DATA. *Yale journal of law and technology*. 2017, Vol.18 P. 201-202. URL:https://digitalcommons.law.yale.edu/cgi/viewcontent.cgi?article=1122&context=yjolt (дата звернення: 07.11.2020)
15. Estabrooks A, Jo T, Japkowicz N, A multiple resampling method for learning from imbalanced data sets. *Computational Intelligence*. 2004, №20, P. 18–36
16. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research.* 2002, № 16, P.321-357.
17. Liu XY, WU J, Zhou ZH. Exploratory undersampling for class imbalanced learning. 2006. P.965–969. *IEEE Transactions on Systems, Man, and Cybernetics.* April 2009. Volume: 39. URL: https://ieeexplore.ieee.org/document/4717268 (дата звернення: 07.11.2020)
18. Joseph Rocca, Baptiste Rocca. Ensemble methods: bagging, boosting and stacking. *Towards data science.*. 2019. №11. P.15.

URL:https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205 (дата звернення: 01.10.2020)

19. Laurikkala J. Improving identification of difficult small classes by balancing class distribution. *Conference on Artificial Intelligence in Medicine in Europe*. University of Tampere. Finland. 2004. P. 63–66. URL: https://link.springer.com/chapter/10.1007%2F3-540-48229-6_9 (дата звернення: 26.10.2020)
20. Wei Q, Dunbrack R.L. The Role of Balanced Training and Testing Data Sets for Binary Classifiers in Bioinformatics. *PLoS ONE*. 2013. № 8. P. 10-12 URL: https://doi.org/10.1371/journal.pone.0067863 (дата звернення: 25.10.2020)
21. Wei Feng, Wenjiang Huang and Jinchang Ren. Class Imbalance Ensemble Learning Based on the Margin Theory. *MDPI Applied Sciences*. 2018. №8. P. 15-23. URL:https://res.mdpi.com/d_attachment/applsci/article_deploy/applsci-08-00815.pdf (дата звернення: 26.10.2020)
22. Collin Ching, Logistic regression theory for practitioners. *Towards data science.* 2020. №2. P.17. URL: https://towardsdatascience.com/the-data-scientists-field-guide-to-logistic-regression-part-1-intuition-97084b11bd68 (дата звернення: 26.10.2020)
23. Benjamin Rogojan. Ensemble Methods to Optimize Machine Learning Model. *Journal of Computer and System Sciences*. 2017. № 3. P 15. URL: https://hub.packtpub.com/ensemble-methods-optimize-machine-learning-models (дата звернення: 01.10.2020)
24. Martin Vojtek, Evzen Kocenda. Credit Scoring Methods. *Article in Finance a Uver journal.* 2006. №53. P. 162-163. URL: https://www.researchgate.net/publication/285873211_Credit_scoring_methods/link/5d4acbb392851cd046a6d72a/download (дата звернення: 26.10.2020)

25. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to Boosting. *Journal of Computer and System Sciences*. 1997. Vol. 55. P.119–139
26. Breiman L. Bagging predictors. Machine Learning. *Statistics Department, University of California*. Berkele. 1996. №24. P.123–140.
27. Wolpert, D.H.: Stacked generalization. Neural Networks. *Complex Systems Group, Theoretical Division, and Center for Non-linear Studies*. 1992. №5 P.241–260.
28. Weiss GM, Provost F The effect of class distribution on classifier learning: An empirical study. *Department of Computer Science*. 2004. № 8. P 27.
29. Huang Te-Ming, Kecman Vojislav, Kopriva Ivica. Kernel Based Algorithms for Mining Huge Data Sets, in Supervised, Semi-supervised, and Unsupervised Learning. *Springer-Verlag*. Berlin. Heidelberg. 2006. P. 260.
30. Ho, Tin Kam. Random Decision Forests. *Proceedings of the 3rd International Conference on Document Analysis and Recognition*. Montreal. 16 August 1995. P. 278–282.
31. Ho, Tin Kam. The Random Subspace Method for Constructing Decision Forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 1998. №2. P. 832–844. URL:https://ieeexplore.ieee.org/document/709601 (дата звернення: 27.10.2020)

ДОДАТОК А. ЛІСТИНГ ПРОГРАМИ

Balancing_dataset.py

```
import pandas as pd
from sklearn.metrics import accuracy_score
from sklearn.linear_model import LogisticRegression
from sklearn.utils import resample
import matplotlib.pyplot as plt
from sklearn.metrics import roc_auc_score

dataset = pd.read_csv("dataset", error_bad_lines=False, sep=',', low_memory=False,
                      index_col=0).astype({'Fails35NoPayIn90': 'int32'})
dataset = dataset.drop("advanceID", axis=1)
df_minority = dataset[dataset['Fails35NoPayIn90'] == 1]
df_majority = dataset[dataset['Fails35NoPayIn90'] == 0]

df_minority_upsampled = resample(df_minority, replace=True,
                                 n_samples=430000, random_state=123)
df_upsampled = pd.concat([df_majority, df_minority_upsampled])
df_upsampled['Fails35NoPayIn90'].hist()
plt.show()
print(df_upsampled['Fails35NoPayIn90'].value_counts())
df_upsampled.to_csv("dataset_balanced")
y = df_upsampled['Fails35NoPayIn90']
X = df_upsampled.drop('Fails35NoPayIn90', axis=1)
clf_1 = LogisticRegression().fit(X, y)
pred = clf_1.predict(X)
pred_pandas = pd.Series(pred)
print(pred_pandas.value_counts())
print(accuracy_score(y, pred))
prob_y_2 = clf_1.predict_proba(X)
prob_y_2 = [p[1] for p in prob_y_2]
print("Roc_Auc_Score" + str(roc_auc_score(y, prob_y_2)))
```

creating_trainig_set.py

```
import pandas as pd
import matplotlib.pyplot as plt
```

```
target_var = pd.read_csv("target.csv", comment='#', error_bad_lines=False, sep=',',
        low_memory=False,index_col=0).astype({'Fails35NoPayIn90': 'int32'})
another_var = pd.read_csv("fixed", comment='#', error_bad_lines=False, sep=',',
        low_memory=False,index_col=0)
print(target_var.info())
print("##########################")
print(another_var.info())
final_df = pd.merge(another_var, target_var, left_on='advanceID', right_on='advanceID', how='left')
print(final_df.info())
print("##########################")
final_df = final_df[final_df['Fails35NoPayIn90']!= '']
print(final_df.describe())
print("##########################")
final_df.dropna(how="any",inplace=True)
final_df=final_df.astype({'Fails35NoPayIn90': 'int32'})
print(final_df.describe())
#final_df.to_csv("dataset")
print(final_df['Fails35NoPayIn90'].value_counts())
final_df['Fails35NoPayIn90'].hist()
plt.show()
```

first_try_prediction.py

```
import pandas as pd
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, roc_curve
from sklearn.metrics import roc_auc_score
from sklearn.model_selection import train_test_split
from matplotlib import pyplot as plt
data_set = pd.read_csv("dataset", error_bad_lines=False, sep=',', low_memory=False,
            index_col=0).astype({'Fails35NoPayIn90': 'int32'})
data_set = data_set.drop("advanceID", axis=1)
y = data_set['Fails35NoPayIn90']
X = data_set.drop('Fails35NoPayIn90', axis=1)
X_train, X_test, y_train, y_test = train_test_split(X, y,
                                    test_size=0.15,
                                    random_state=42)
clf_0 = LogisticRegression().fit(X_train, y_train)
pred = clf_0.predict(X_test)
print(accuracy_score(pred, y_test))
pred_pandas = pd.Series(pred)
print(pred_pandas.value_counts())
```

```
prob_y_2 = clf_0.predict_proba(X_test)
prob_y_2 = [p[1] for p in prob_y_2]
ns_probs = [0 for _ in range(len(y_test))]
print(roc_auc_score(y_test,prob_y_2))
# calculate scores
ns_auc = roc_auc_score(y_test, ns_probs)
lr_auc = roc_auc_score(y_test, prob_y_2)
# summarize scores
print('No Skill: ROC AUC=%.3f' % (ns_auc))
print('Logistic: ROC AUC=%.3f' % (lr_auc))
# calculate roc curves
ns_fpr, ns_tpr, _ = roc_curve(y_test, ns_probs)
lr_fpr, lr_tpr, _ = roc_curve(y_test, prob_y_2)
# plot the roc curve for the model
plt.plot(ns_fpr, ns_tpr, linestyle='--', label='No Skill')
plt.plot(lr_fpr, lr_tpr, marker='.', label='Logistic')
# axis labels
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
# show the legend
plt.legend()
# show the plot
plt.show()
```

log_reg_balancing.py

```
import pandas as pd
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, roc_curve
from sklearn.metrics import roc_auc_score
from sklearn.model_selection import train_test_split
from matplotlib import pyplot as plt
from sklearn.utils import resample

data_set = pd.read_csv("dataset", error_bad_lines=False, sep=',', low_memory=False,
                       index_col=0).astype({'Fails35NoPayIn90': 'int32'})
data_set = data_set.drop("advanceID", axis=1)
y = data_set['Fails35NoPayIn90']
X = data_set.drop('Fails35NoPayIn90', axis=1)
df_minority = data_set[data_set['Fails35NoPayIn90'] == 1]
df_majority = data_set[data_set['Fails35NoPayIn90'] == 0]
```

```
df_minority_upsampled = resample(df_minority, replace=True,
                        n_samples=430000, random_state=123)
df_upsampled = pd.concat([df_majority, df_minority_upsampled])

df_upsampled['Fails35NoPayIn90'].hist()
y = df_upsampled['Fails35NoPayIn90']
X = df_upsampled.drop('Fails35NoPayIn90', axis=1)
plt.show()
X_train, X_test, y_train, y_test = train_test_split(X, y,
                                    test_size=0.15,
                                    random_state=42)
clf_0 = LogisticRegression().fit(X_train, y_train)
pred = clf_0.predict(X_test)
print(accuracy_score(pred, y_test))
pred_pandas = pd.Series(pred)
print(pred_pandas.value_counts())
prob_y_2 = clf_0.predict_proba(X_test)
prob_y_2 = [p[1] for p in prob_y_2]
ns_probs = [0 for _ in range(len(y_test))]
print(roc_auc_score(y_test,prob_y_2))
# calculate scores
ns_auc = roc_auc_score(y_test, ns_probs)
lr_auc = roc_auc_score(y_test, prob_y_2)
# summarize scores
print('No Skill: ROC AUC=%.3f' % (ns_auc))
print('Logistic: ROC AUC=%.3f' % (lr_auc))
# calculate roc curves
ns_fpr, ns_tpr, _ = roc_curve(y_test, ns_probs)
lr_fpr, lr_tpr, _ = roc_curve(y_test, prob_y_2)
# plot the roc curve for the model
plt.plot(ns_fpr, ns_tpr, linestyle='--', label='No Skill')
plt.plot(lr_fpr, lr_tpr, marker='.', label='Logistic')
# axis labels
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
# show the legend
plt.legend()
# show the plot
plt.show()
```