Міністерство освіти і науки України
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «КИЄВО-МОГИЛЯНСЬКА АКАДЕМІЯ»
Кафедра мультимедійних технологій факультету інформатики

Аналіз технологій машинного навчання на прикладі успішності стартапів

Текстова частина до курсової роботи
за спеціальністю «Інженерія програмного забезпечення» 121

Керівник курсової роботи
Доцент, кандидат фізико-математичних наук
Жежерун О. П.

_____
(підпис)
“____” _____ 2020 р.
Виконала студентка 3-го курсу
Хоменець В.С.
“____” _____ 2020 р.

Київ 2020

Календарний план виконання курсової роботи

Тема: Аналіз технологій машинного навчання на прикладі успішності стартапів

Календарний план виконання роботи:

№ п/п	Назва етапу курсової роботи	Термін виконання етапу	Примітка
1.	Розгляд проблеми та збір потрібної інформації	Вересень-листопад 2019р.	
2.	Аналіз технологій	Грудня-лютий 2020р.	
3.	Розробка архітектури програми та інтерфейсу	Лютий-березень 2020р.	
4.	Розробка частини машинного навчання	Березень-квітень 2020р.	
5.	Написання частин front-end і back-end	Квітень 2020р.	
6.	Перегляд праці науковим керівником	Травень 2020р.	
7.	Створення презентації роботи	Травень 2020р.	
8.	Презентація роботи		

Студент Хоменець В.С _____

Керівник Жежерун О.П _____

“_____” _____ 2020 р.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ

1. DS - Data Science
2. ML - Machine Learning
3. AI - Artificial Intelligence
4. API - Application Programming Interface
5. SQL - Structured Query Language
6. SAS - Statistical Analysis Software

АНОТАЦІЯ

В даній роботі розглянуто технології Data Science і Machine Learning. Для їхнього аналізу було створено готовий для користування продукт, метою якого є передбачення успішності стартапів. Це надає змогу керівникам проектів знати поточний шлях розвитку продукту і корегувати оновлення відносно нього.

Було використано передові технології для того, щоб зробити застосунок якомога адаптивнішим до змін ринку і щоб всі вимоги користувача були задоволені якнайбільше. Для розробки front-end було застосовано jQuery, back-end - Node.js, машинного навчання - Python з його бібліотеками.

ВСТУП

В сучасному світі повз нас проходить безмежна кількість інформації кожної секунди. Нам з цього потрібно брати стільки користі, наскільки це можливо. Саме тому у світі з’явилися напрямки Data Science, який вказує як можна аналізувати необроблену інформацію, та Machine Learning, який на основі великих датасетів дає змогу створити програму, яку не можливо створити стандартним підходом.

Для аналізу вище згаданих технологій потрібно було створити корисний для людства продукт. Після довгих днів міркувань помічено проблему, яка виникає у кожного власника застосунку. Ця проблема полягала у тому, що керівник проекту не до кінця розуміє в якому шляху на певний момент часу рухається його компанія. Було прийнято рішення створити продукт, який допоможе розв’язати таку проблему.

У чому полягає основний сценарій роботи продукту? Для того, щоб визначити шлях розвитку проекту, у застосунку для початку реєструється керівник проекту. Він створює проект і додає його працівників. Керівник в певний момент часу надсилає опитування працівникам. Після проходження працівниками опитування менеджер може побачити відсоток ймовірності успіху і прийняти потрібні рішення.

Які питання отримують співробітники? Було проаналізовано історії різних продуктів і що саме впливало на їх успіх. Рішення після аналізу полягало у тому, що потрібно поставити три запитання до працівників:

- Чи достатньо був проаналізований вплив оновлення на етапі його розробки?
- Чи достатньо була проаналізований складність оновлення на етапі його розробки?
- Чи оновлення ведуть до покращення ситуації в компанії?

Як часто потрібно проводити опитування персоналу? Коли менеджер перестає бути впевнений в успішності розвитку застосунку або після

випуску оновлення, варто переконатися чи все йде за планом. Проте це все ж таки вирішувати самому керівнику.

Робота складається з чотирьох частин.

Перша частина присвячена аналізу технологій, які використовувалися для розробки продукту.

У другів частині аналізується підбір запитань для опитування працівників.

Третя частина описує використання машинного навчання для передбачення успішності проекту.

Четверта частина присвячена опису практичного завдання.

АНАЛІЗ НАПРЯМКІВ DATA SCIENCE І MACHINE LEARNING

1.1 Загальні відомості про Data Science

Data Science - це область дослідження, яка передбачає отримання розуміння поведінок із величезної кількості даних за допомогою різних наукових методів, алгоритмів та процесів. Це допомагає виявити приховані зразки поведінки з необроблених даних. Термін Data Science з'явився через еволюцію математичної статистики, аналізу даних та великих даних.

Data Science - це міждисциплінарна сфера, яка дозволяє отримувати знання зі структурованих чи неструктурованих даних. Наука даних дозволяє перетворити бізнес-проблему на дослідницький проект, а потім перетворити її назад у практичне рішення.

Використання технології аналізу даних (Data Analytics) має суттєві переваги а саме:

- Дані - це нафта для сучасного світу. За допомогою правильних інструментів, технологій, алгоритмів ми можемо використовувати дані та перетворювати їх у відмінні переваги для бізнесу
- Data Science може допомогти вам виявити шахрайство за допомогою сучасних алгоритмів машинного навчання
- Дана технологія допомагає вам запобігти будь-які значні грошові втрати
- Дозволяє формувати інтелекту в машинному навчанні
- Ви можете провести аналіз настроїв, щоб оцінити лояльність клієнта
- Це дає змогу приймати кращі та швидші рішення
- Допомагає рекомендувати правильний товар правильному клієнту для розширення бізнесу

Data Science складається з наступних компонентів:

- Статистика. Є найбільш критичною одиницею в науці даних. Це метод чи наука збирання та аналізу чисельних даних у великих кількостях для отримання корисної інформації.
- Візуалізація. Техніка візуалізації допомагає отримати доступ до величезної кількості даних у легко зрозумілих та засвоюваних візуальних зображеннях.
- Машинне навчання. Вивчає побудову та вивчення алгоритмів, які вчяться робити прогнози щодо непередбачених / майбутніх даних.
- Глибоке навчання (Deep Learning). Метод глибокого навчання - це нове дослідження машинного навчання, де алгоритм вибирає модель аналізу, яку слід використовувати у тій чи іншій ситуації.

1.2 Процес Data Science

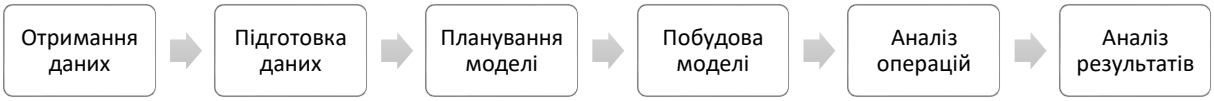

Рисунок 1. Процес Data Science

1.2.1 Отримання даних

Даний крок передбачає отримання даних з усіх виявлених внутрішніх та зовнішніх джерел, що допомагає відповісти на поставлені бізнес-питання.

Потенціальні джерела даних:

- Журнали від веб-серверів
- Дані, зібрані з соціальних медіа
- Набори даних перепису
- Дані передані з інтернет-джерел за допомогою API

1.2.2 Підготовка даних

У даних може бути безліч невідповідностей, таких як відсутнє значення, порожні стовпці, неправильний формат даних, який потрібно відредагувати. Перед моделюванням потрібно обробити, дослідити та перевірити стан даних. Чим чистіші вхідні дані, тим достовірнішими є прогнози.

1.2.3 Планування моделі

На цьому етапі потрібно визначити спосіб і техніку аналізу, щоб скласти співвідношення між вхідними змінними. Планування моделі виконується за допомогою різних статистичних формул та засобів візуалізації. Служби аналізу SQL, R та SAS / Access – приклади деяких інструментів, що використовуються для цієї мети.

1.2.4 Побудова моделі

На цьому кроці починається власне процес побудови моделі. Тут Data Scientist розподіляє набори даних для навчання та тестування. Такі методи, як об'єднання, класифікація та кластеризація, застосовуються до набору даних для тестування. Розроблена модель тестується на наборі даних, що призначені для тестування.

1.2.5 Аналіз операцій

На цьому етапі надається остаточна базова модель із звітами, кодом та технічними документами. Модель розміщена у виробничому середовищі в режимі реального часу після ретельного тестування.

1.2.6 Аналіз результатів

На цьому етапі ключові висновки повідомляються всім зацікавленим сторонам. Це допомагає визначити, чи результати проекту є успішними чи невдалими на основі вхідних даних моделі.

1.3 Загальні відомості про Machine Learning

Машинне навчання (Machine Learning) - це складова частина штучного інтелекту (Artificial Intelligence), яка надає системам можливість технічно досліджувати інформацію та вдосконалюватись без явного запрограмування. Машинне навчання спрямоване на розробку мов програмування, які можуть отримувати доступ до статистики та використовувати її для аналітичних та наукових цілей. Основна мета машинного навчання - дозволити комп'ютерним системам регулярно аналізувати без втручання людини чи допомоги та змінювати рухи з цієї причини.

1.4 Алгоритми машинного навчання

Алгоритми машинного навчання поділяються класифікуються як керовані та некеровані.

Керовані алгоритми машинного навчання застосовують те, що було вивчено раніше, до нових даних, використовуючи помічені приклади для прогнозування майбутніх подій. Починаючи з аналізу відомого навчального набору даних, алгоритм навчання виробляє певну функцію для прогнозування вихідних значень. Система здатна забезпечити цілі для будь-якого нового входу після достатньої підготовки. Алгоритм навчання може

також порівнювати його результат з правильним, призначеним результатом та знаходити помилки, щоб відповідно модифікувати модель.

Некеровані алгоритми машинного навчання у свою чергу використовуються, коли інформація, що використовується для тренування, не є ні класифікованою, ні маркованою. Без нагляду навчання вивчає як системи можуть зробити висновок про функцію опису прихованої структури з не маркованих даних. Система не знаходить правильного виводу, але вона досліджує дані і може робити висновки з наборів даних для опису прихованих структур з не маркованих даних.

ПІДБІР КЛЮЧОВИХ ФАКТОРІВ ВПЛИВУ НА УСПІШНІСТЬ ПРОЕКТУ

2.1 Пошук інформації

Одним із найважчих завдань під час створення застосунку на базі Машинного навчання є пошук інформації. Чим більше інформації зібрано, тим правдивіше будуть передбачення.

Для того, щоб створити застосунок передбачення успішності проекту, потрібно було зібрати багато даних від їх працівників. Це надзвичайно важка робота, яка потребує не лише часу, але й відповідне коло знайомств. Оскільки метою проекту є проаналізувати технологію розробки такого продукту, датасет був придуманий не з дійсної інформації. Проте, після виконаного практичного завдання, все що залишається зробити, щоб продукт став комерційним – це накопичити дійсну інформацію.

2.2 Побудова моделі

Головними запитаннями, які будуть задаватися працівникам під час кожного опитування є:

- Чи достатньо був проаналізований вплив оновлення на етапі його розробки?
- Чи достатньо була проаналізований складність оновлення на етапі його розробки?
- Чи оновлення ведуть до покращення ситуації в компанії?

Аналіз впливу оновлення є одним із ключових факторів успішності застосунку. Часто бувають ситуації, коли новий функціонал робить шкоди продукту, або його розробка забирає дуже багато зусиль команди, що себе не виправдовує. Проте оновлення варто робити періодично, щоб підтримувати продукт у відповідності до вимог ринку.

Аналіз складності оновлення теж дуже важливий фактор. В період розробки нового функціоналу часто виникає проблема, коли терміни його завершення порушуються. Це призводить до погіршення якості застосунку і фінансових втрат.

Неможливо розробляти передбачення успішності проекту без запитання чи оновлення ведуть до покращення компанії.

ВИКОРИСТАННЯ МАШИННОГО НАВЧАННЯ ДЛЯ ПЕРЕДБАЧЕННЯ УСПІШНОСТІ ПРОЕКТУ

Якщо потрібно виконати задачу за допомогою технологій машинного навчання одразу спадає на думку використати або Python, або R. Оскільки Python легше освоїти і навики його володінням можна застосувати і у інших сферах було обрано саме його.

3.1 Мова програмування Python

3.1.1 Визначення

Python (найчастіше вживане прочитання — «Па́йтон», запозичено назву з британського шоу Монті Пайтон) — інтерпретована об'єктно-орієнтована мова програмування високого рівня зі строгою динамічною типізацією. Розроблена в 1990 році Гвідо ван Россумом.

3.1.2 Загальний опис

Багато коду для створення обчислень різних даних вже написано на Python. Для того, щоб їх використовувати потрібно підключати різні бібліотеки. Зокрема, під час виконання практичної частини було використано numpy – базовий контейнер, який містить функції і об’єкти для наукових обчислень, pandas – бібліотека для аналізування даних, sklearn – бібліотека, за допомогою якої реалізується технології машинного навчання (лінійна регресія, випадковий ліс).

Python також знайшов своє застосування у розробці серверної частини. Існують такі фрейморки як Django, Pyramid.

3.2 Передбачення успіху

3.2.1 Створення датасету

Уся інформація зберігалася у базі даних, яка було розроблена для даного проекту (див. рисунок 2).

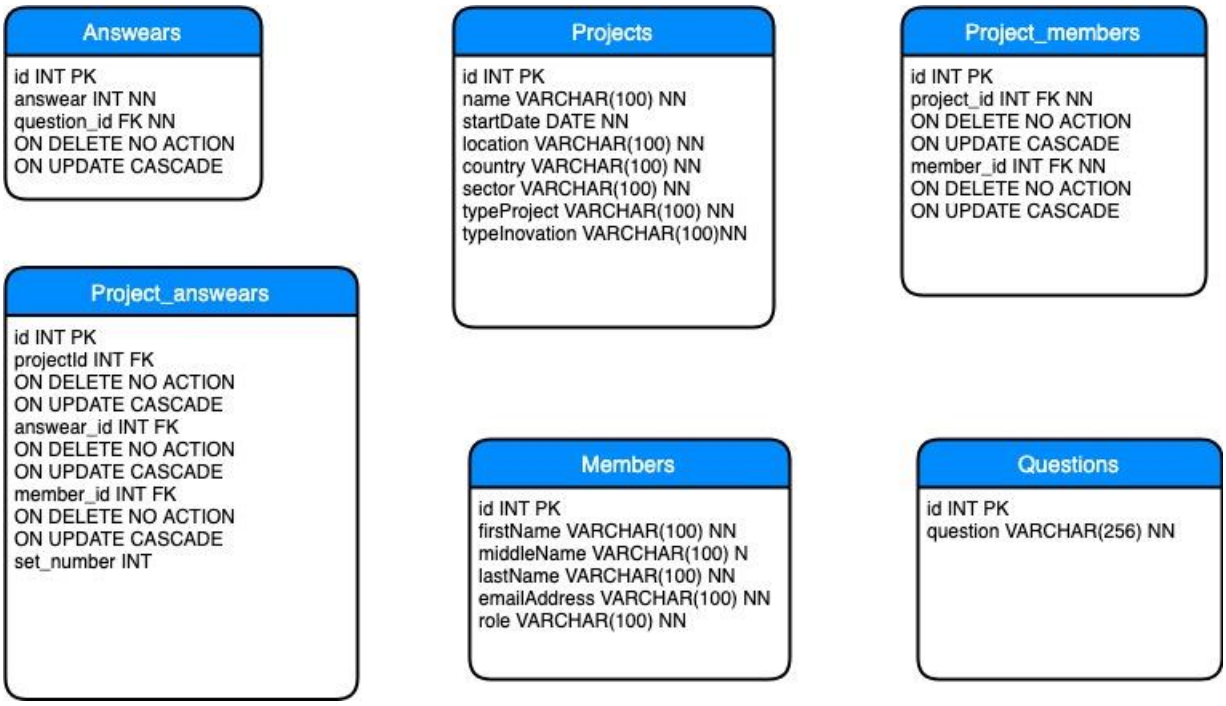

Рисунок 2. Схема бази даних застосунку

Спочатку потрібно було сформувати датасет в структуру даних. Для цього потрібно стягнути усю потрібну інформацію з бази даних. Метод для встановлення зв’язку з базою даних описано нижче.

```
def get_db_connection():
    return mysql.connector.connect(
      host="localhost",
      user="successProjectDeveloper",
      passwd="Success1_",
      auth_plugin='mysql_native_password'
    )
```

Формування датасету відбувається у після виконання декількох трансакцій з бази даних декілька.

- Отримуємо кількість проектів

```
query = ("select count(*) from Project")
cursor.execute(query)
row_count = cursor.fetchall()[0][0]
```

- Отримуємо всі запитання

```
query = ("select * from Question")
cursor.execute(query)
questions = []
for (_, question) in cursor:
    questions.append(question)
```

- Формуємо структуру даних висотою, яка дорівнює кількості проектів і шириною, яка дорівнює кількості запитань

```
dataset = np.zeros((row_count, len(questions)), dtype=np.int)
```

- Створюємо структуру answers, у якій будуть у відповідності зберігатися відповіді і запитання

```
query = ("select count(*) from Answers")
cursor.execute(query)
answers_count = cursor.fetchall()[0][0]
answers = np.zeros((answers_count, 2), dtype=np.int16)

query = ("select * from Answers")
cursor.execute(query)
for (id, answer, question_id) in cursor:
    question_id = float(str(question_id))
    answer = str(answer)
    if answer == 'None':
        answer = 0
    else:
        answer = float(answer)
    answers[id - 1, 0] = question_id
    answers[id - 1, 1] = answer
```

- Записуємо у структуру dataset відповідно відповіді на три запитання до кожного проекту

```
query = ("select * from Project_Answers")
cursor.execute(query)
for (_, project_id, answer_id, _, _) in cursor:
    project_id = int(str(project_id))
    answer_id = int(str(answer_id))
    dataset[project_id - 1, answers[answer_id - 1][0] - 1] = answers[answer_id - 1][1]
```

3.2.2 Передбачення за допомогою лінійної регресії

Для початку потрібно створити об’єкт класу LinearRegression і пропустити через нього інформацію для тренування.

```
def process_dataset(questions, dataset):
    X = dataset[questions[:len(questions) - 1]]
    Y = dataset[questions[-1]]
    regr = linear_model.LinearRegression()
    regr.fit(X, Y)
    return regr
```

Потім залишається лише викликати метод predict класу LinearRegression з передавання параметра масива, у якому містяться значення середніх арифметичних відповідей працівників на кожне запитання відповідно.

3.3 Проблеми розробки

Під час створення продукту було обрано використовувати Node.js для написання back-end частини застосунку. Це призвело до ускладнень під час її інтеграції з кодом, призначеним для надавання передбачення і написаним на мові Python.

- Чи достатньо був проаналізований вплив нового оновлення на етапі його розробки?
- Чи достатньо була проаналізований складність нового оновлення на етапі його розробки?
- Чи оновлення ведуть до покращення ситуації в компанії?

ОПИС ПРАКТИЧНОГО ЗАВДАННЯ

4.1 Головне меню керівника проекту

Після автентифікації керівник проекту бачить головне вікно. На ньому у нього є можливість перейти на:

- Перегляд проектів
- Перегляд запитань
- Перегляд команд працівників
- Перегляд працівників
- Перегляд відповідей на опитування

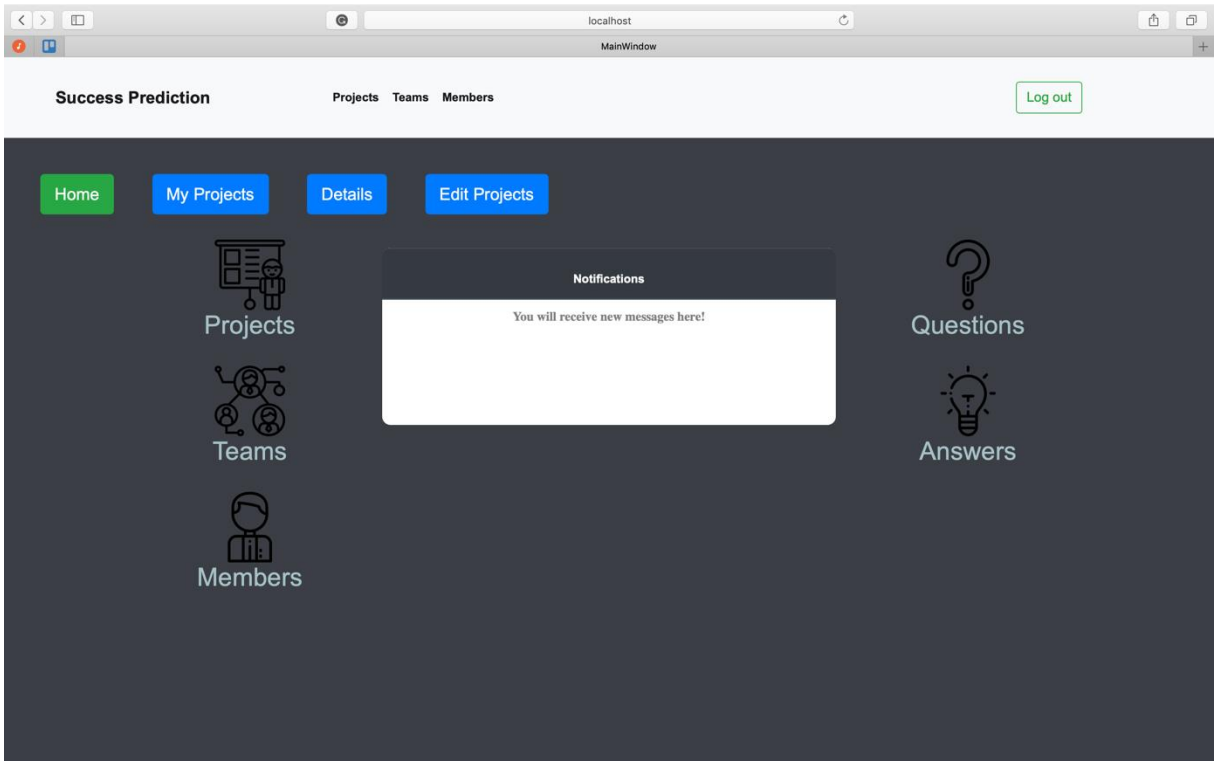

Рисунок 3. Головне вікно застосунку

4.2 Створення і редагування проекту

Керівник проекту може створювати і редагувати проекти. Для створення йому потрібно ввести таку інформацію:

- Ім’я проекту
- Початкова дата проекту
- Континент, на якому проект буде запущений
- Галузь проекту
- Тип проекту
- Тип інновації, яку несе проект

Також він може залучати працівників до проекту. Після цього існує опція надіслати опитування.

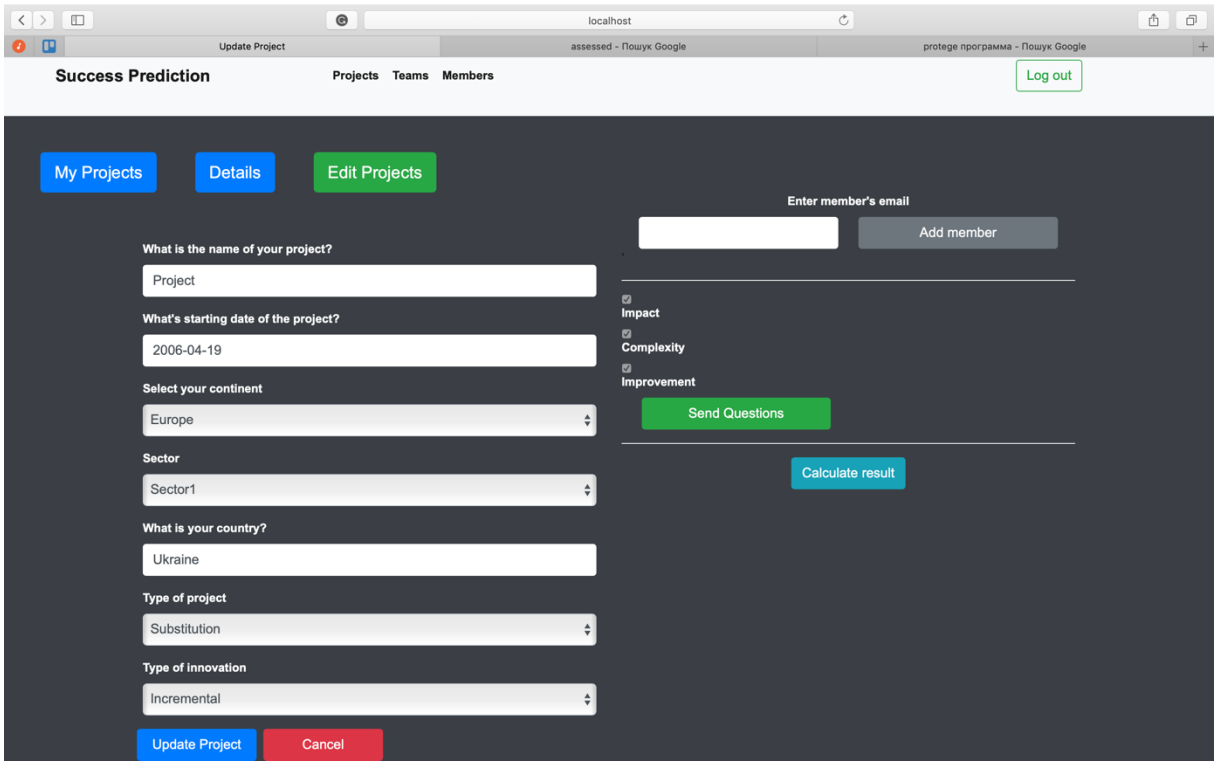

Рисунок 4. Опитування

4.3 Перегляд проектів

Менеджер проектів може їх переглядати у таблиці, в якій описана детальна їхня інформація.

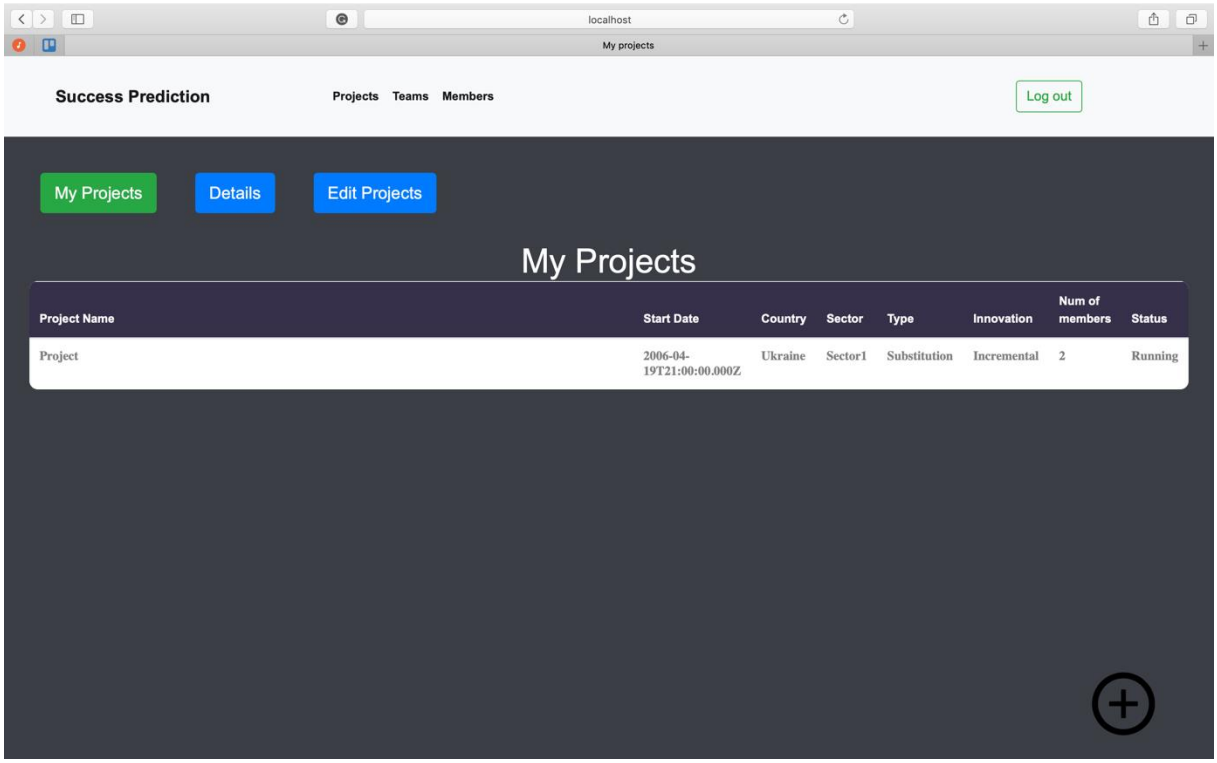

Рисунок 5. Таблиця проектів

4.4 Головне меню працівника проекту

Працівник отримує сповіщення про те, що йому потрібно пройти опитування і тоді у нього з’являється таблиця, у якій вказано всі непройдені проекти, де потрібна його участь.

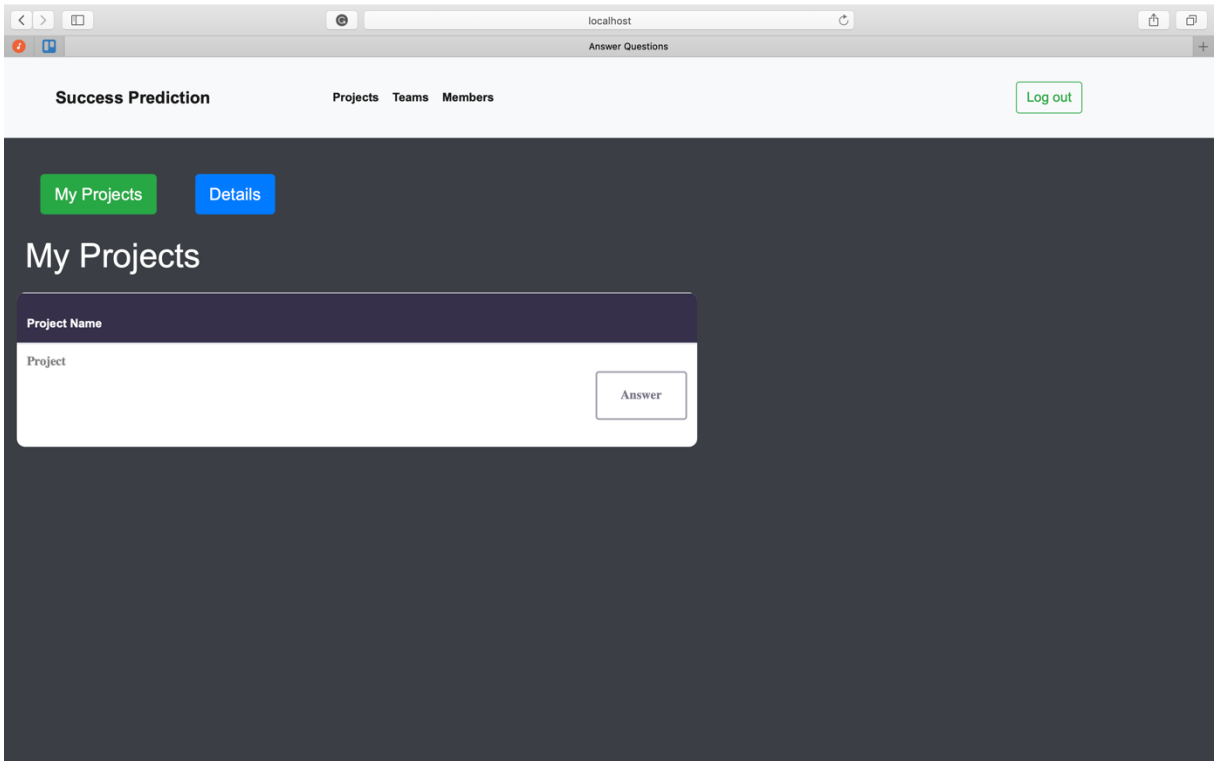

Рисунок 6. Проекти, в яких необхідно пройти опитування

4.5 Опитувальник

Опитувальник має три запитання. Працівник після надання відповідей на них натискає кнопку збереження опитування.

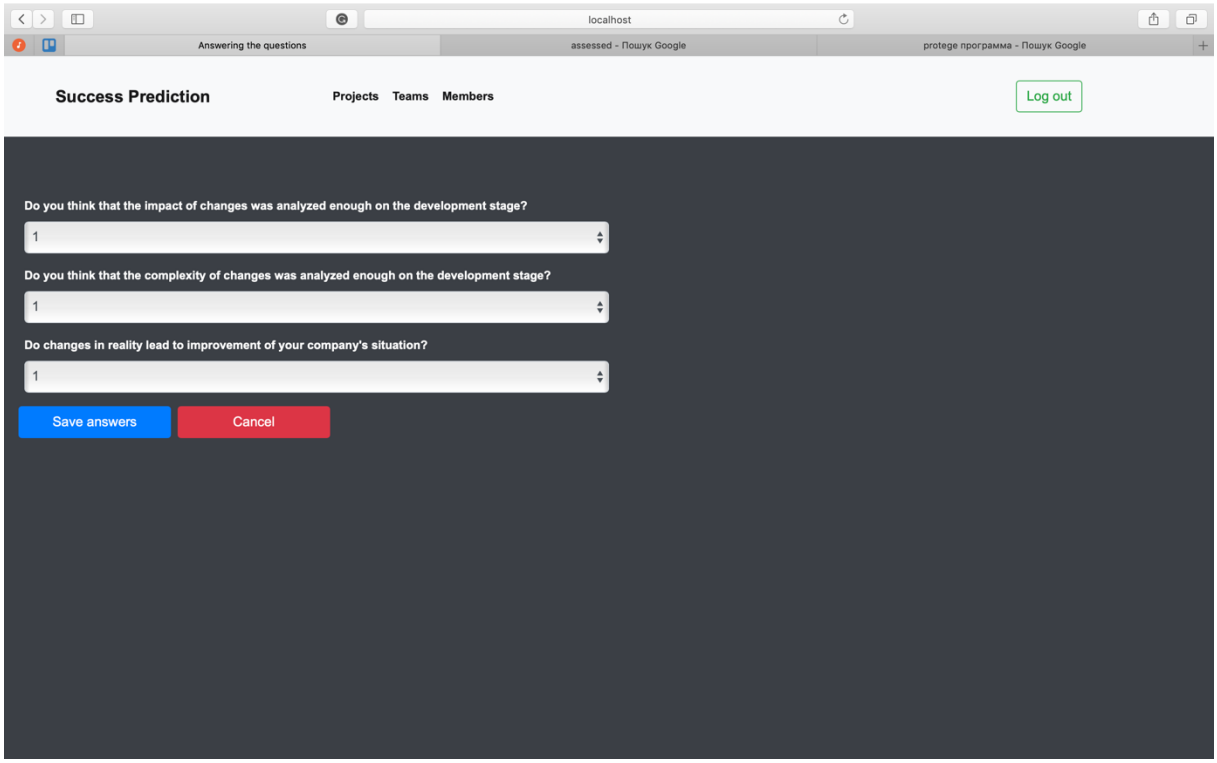

Рисунок 7. Запитання по проекту

4.6 Передбачення успішності проекту

Керівник проекту після проходження опитування працівниками, переходить за покликанням порахувати результат і бачить відсоток ймовірності успіху його проекту на даний момент.

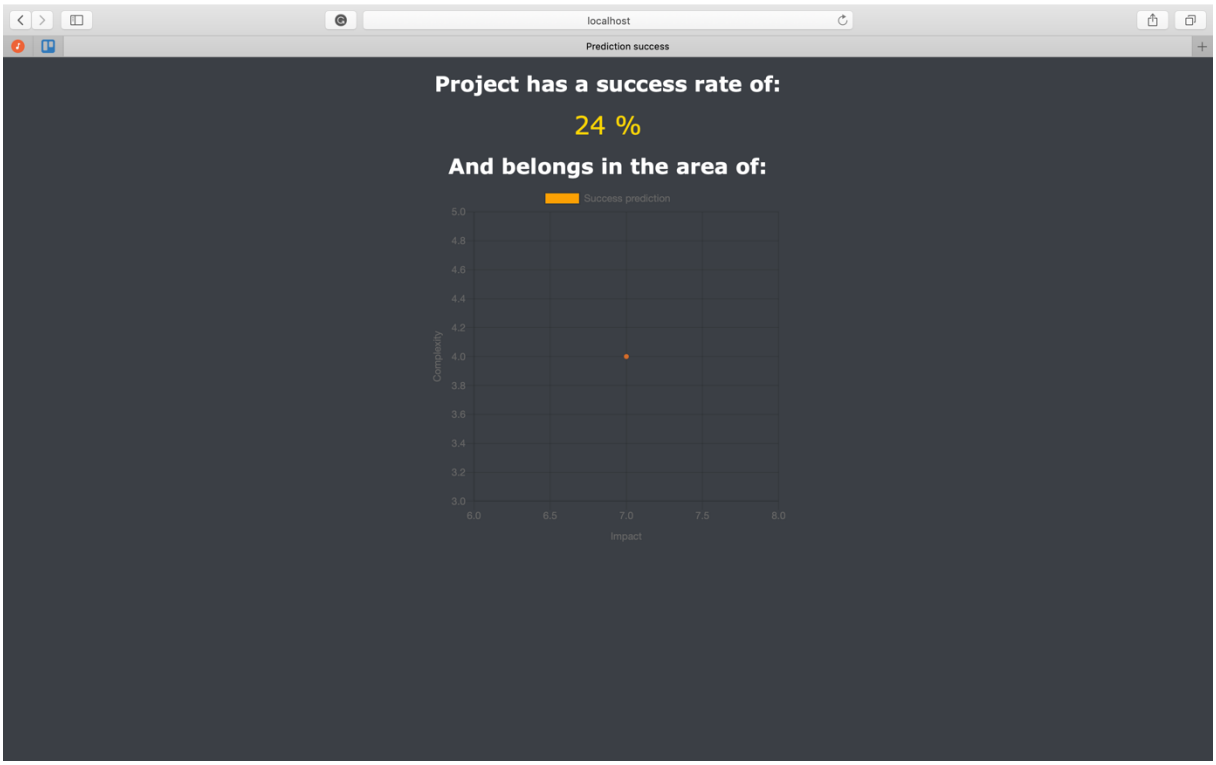

Рисунок 8. Відсоток ймовірності успіху проекту

ВИСНОВОК

Під час виконання даної роботи було розглянуто багато сучасних технологій і створено продукт, який корисний для багатьох керівників проектів. Не залишилося сумніву, що такі технології як Data Science і Machine Learning дають змогу створити продукти, про які раніше ніхто б і не зміг подумати.

Було поглиблено знання у сфері стартапів, ймовірності їх успішності або невдачі. Тим самим проаналізовано основні фактори успіху проекту. Це дасть можливість у майбутньому створювати більше успішних застосунків, ніж провальних.

Не менш важливим є і покращення навичок у таких технологіях як: jQuery, Node.js і Python.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Data Science - https://www.guru99.com/data-science-tutorial.html
2. Data Science - https://www.edureka.co/blog/what-is-data-science/
3. Machine Learning - https://emerj.com/ai-glossary-terms/what-is-machine-learning/
4. Machine Learning - https://expertsystem.com/machine-learning-definition/
5. "Основи Data Science та Big Data. Python та наука про дані" Деві Сілен, Арно Мейсман, Мохамед Алі
6. "Введення до машинного навчання за допомогою Python" Андреас Мюлер, Сара Гвідо
7. "Самонавчаючі системи" С.І. Ніколаєнко, А.Л. Тулупьєв
8. "Математичні основи теорії машинного навчання та прогнозування" В.В. Вьюгін
9. "Машинне навчання" Брінк Х., Річардс Дж., Феверолф М.
10. "Python для складних задач. Наука про дані та машинне навчання" Плас Дж.В.