МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ЧОРНОМОРСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ПЕТРА МОГИЛИ

Овчар Катерина Віталіївна

УДК 303.724.32.039.3

СКОРІНГОВА СИСТЕМА ЗАКЛЮЧЕННЯ КОНТРАКТІВ В ІТ-СФЕРІ

Спеціальність 123 – Комп’ютерна інженерія

Автореферат
магістерської роботи
на здобуття кваліфікації магістра з комп’ютерної інженерії

Миколаїв – 2021

Робота виконана у Чорноморському національному університеті ім. Петра Могили.

Керівник: **Дворник Ольга Василівна,**
ЧНУ ім. Петра Могили,
канд. фіз.-мат. наук,
доцент

Рецензент: **Кондратенко Юрій Пантелійович,**
ЧНУ ім. Петра Могили,
завідувач кафедри інтелектуальних
інформаційних систем,
доктор технічних наук,
професор

Захист відбудеться «23» лютого 2021 року на засіданні Екзаменаційної комісії в ЧНУ ім. Петра Могили, ауд. 2-406.

З магістерською роботою можна ознайомитись на сайті ЧНУ ім. Петра Могили за посиланням http://chmnu.edu.ua.

Автореферат оприлюднений « 22 » лютого 2021 р.

Секретар
екзаменаційної комісії,
кандидат фіз.-мат. наук, доцент С. В. Пузирьов

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми. Станом на 2020 рік, Україна є одним з провідних ринків IT-аутсорсингу в світі. Тільки за 2017 рік, в нашій країні було засновано 3000 компаній, що надають послуги з розробки програмного забезпечення іноземним установам та стартапам. Для таких компаній, критично важливим є процес лідогенерації (пошуку потенційних клієнтів та налагодження зв'язку з ними). Пошук клієнта для заключення одного контракту може коштувати компанії від 700 доларів США, тому аутсорсингові IT-компанії постійно шукають способи скорочення вартості процесу лідогенерації при збереженні або покращенні якості заключених контрактів.

При відборі потенційних клієнтів, лідогенератори враховують ряд факторів, такі як часовий пояс розташування клієнта, наявність конкретного запиту та його детальний опис, рейтинг потенційного замовника на платформах Upwork, Clutch.co, LinkedIn та інших, об'єм коштів, які клієнт вже витратив на послуги розробників. Втім, аналіз запитів вручну потребує значних затрат часу з боку фахівців, і не виключає помилок через людський фактор.

Іншою проблемою, що також суттєво впливає на вірогідність заключення контрактів з потенційними клієнтами в IT, є відсутність швидкого та простого доступу до записів переговорів з клієнтами. Дуже часто при формуванні естімейтів (оцінки вартості та тривалості робіт), необхідно використовувати не лише наявну проектну документацію, але й звертатися до записів проведених переговорів з клієнтами, оскільки вони можуть містити інформацію, необхідну для точної оцінки. Зберігання записів у форматі аудіо- чи відеозаписів є досить незручним з двох причин. По-перше, знайти необхідну інформацію у даних в текстовому форматі набагато швидше, ніж переглядати відеозапис дзвінка. По-друге, рівень

володіння англійською мовою не завжди дозволяє технічним фахівцям у повній мірі зрозуміти вимоги замовника, прослуховуючи запис розмови з ним.

На даний час відомі наукові дослідження та технологічні рішення імплементації методів прогнозування продажів та ціноутворення за допомогою машинного навчання для загального використання, або для конкретних індустрій (таких як роздрібна торгівля, індустрія моди, туризм, тощо). Втім, в наукових джерелах майже не згадується подібних розроблень для сфери IT-аутсорсингу, що і обумовлює актуальність даної роботи.

Мета: розроблення апаратно-програмного комплексу оптимізації процесу заключення контрактів для аутсорсингових IT-компаній шляхом регресійного аналізу наявних даних про поточних клієнтів компанії, а також спектрального аналізу голосів учасників переговорів.

Для досягнення поставленої мети необхідно вирішити такі **завдання**:

– з аналітичного огляду літератури та патентної інформації сформулювати завдання дослідження та розроблення;

– проаналізувати алгоритми автоматичного розпізнавання мовця та регресійного аналізу, обґрунтувати вибір ефективних для розв’язання поставлених задач;

– створити датасет для регресійного аналізу, що складатиметься з реальних даних про клієнтів IT компанії;

– обрати оптимальні ознаки для регресійного аналізу;

– розробити блок-схему алгоритму роботи веб-застосунку, який поєднуватиме функціонал обох алгоритмів;

– підібрати мову програмування для реалізації алгоритмів, та реалізувати їх з використанням обраних засобів;

– здійснити тестування реалізованих алгоритмів та розробленого веб-застосунку;

– розробити питання з цивільного захисту та охорони праці на підприємстві, дані якого використовуються для аналізу.

Об’єкт: методи та засоби цифрового перетворення аналогових акустичних сигналів, зокрема голосів людей, і подальшої їх обробки; методи розпізнавання мовця засобами мови програмування Python; методи навчання регресійної моделі; метод статистичного аналізу.

Предмет: програмний комплекс для визначення ймовірності заключення контракту з потенційним клієнтом на основі даних про поточних клієнтів, введених фахівцем з лідогенерації, та для підвищення конверсії потенційних клієнтів завдяки розпізнаванню аудіозаписів переговорів аутсорсингової ІТ-компанії, розташованої в місті Миколаєві.

Використані методи: лінійна регресія, експеримент, статистичні методи, тестування.

Практичне значення одержаних результатів: результати даної роботи можуть бути використані в аутсорсингових ІТ-компаніях України та інших країн для прогнозування продажів та зниження вартості процесу лідогенерації.

Апробація результатів магістерської роботи відбулася під час Всеукраїнської науково-практичної конференції молодих вчених, аспірантів і студентів «Інтелектуальні інформаційні системи» 2021 р.

Публікації. За результатами магістерської роботи опубліковані тези доповідей [1].

Структура та обсяг роботи. Магістерська робота складається з анотації на 2 сторінках, вступу, чотирьох розділів, висновків, переліку джерел посилання з 31 найменуваня, 5 додатків на 7 сторінках, спеціальної частини з охорони праці та безпеки життєдіяльності. Основна частина роботи становить 71 сторінку, серед яких 20 рисунків та 15 таблиць.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У **вступі** обґрунтовано актуальність теми магістерської роботи – необхідність оптимізації продажів в IT-аутсорсингу як одній з ключових галузей промисловості України. Наведено мету, об’єкт та предмет дослідження, а також завдання, поставлені для досягнення мети. Вказано методи дослідження, використані при виконанні магістерської роботи (лінійна регресія, статистичні методи, експеримент, та тестування). Описано практичне значення роботи. Наведено інформацію про конференцію, в рамках якої відбулася апробація результатів роботи, та тези, опубліковані за результатами матеріалів роботи.

У **першому розділі** магістерської роботи «**Огляд рішень автоматичного скорингу клієнтів та методів розпізнавання мовця**» проведено огляд наявних методів вирішення основних задач, що покладено в основу розроблення програмного комплексу – автоматичного скорингу клієнтів та методів розпізнавання мовця.

Розглянуто методи автоматичного скорингу клієнтів, засновані на методах лінійної регресії, алгоритму Random Forest, та комбінації алгоритмів XGBoost та LightGBM. З порівняльної характеристики цих алгоритмів, зроблено ряд висновків:

– Точність роботи моделі більш суттєво залежить від обраного алгоритму, ніж від розміру навчального датасету.

– Алгоритми, що базуються на роботі з деревами прийняття рішень, є найбільш ефективними та широко використовуваними при кваліфікації клієнтів.

– Оптимальна кількість записів для навчання та тестування моделі – 400±50.

– Оптимальна кількість параметрів, за якими класифікуються клієнти – 20±5.

– Найбільш ефективної роботи методів машинного навчання можна досягти за умови, що частка даних про клієнтів, які в подальшому скористались послугами компанії приблизно дорівнюватиме частці даних про клієнтів, які не стали замовниками.

– Автоматизація класифікації потенційних клієнтів має виконуватись індивідуально для кожного підприємства, оскільки набір даних та оцінюваних параметрів залежить від галузі, розташування цільової аудиторії, тощо.

Також розглянуто різні алгоритми та методи автоматичного визначення мовця. Встановлено, що найбільш точними є алгоритми розпізнавання мовця, що використовують d-вектори як форму представлення даних про людський голос.

У **другому розділі** магістерської роботи **«Математичні методи та проектування системи»** наведено детальний опис алгоритму d-векторів, що використаний в розробленій системі для автоматичного розпізнавання мовця, та регресійних методів, що використані для скорингу клієнтів IT компанії. Наведено математичні методи, використані в даних алгоритмах, та інформацію про основні принципи їхньої роботи.

В розробленій системі використано кожен із згаданих методів регресійного аналізу (множинна регресія, поліноміальна регресія, SVM-регресія, дерева прийняття рішень, та алгоритм Random Forest), оскільки для вирішення задачі скорингу клієнтів в IT компанії доцільно обрати алгоритм, що показуватиме найбільшу точність для конкретного набору ознак. При цьому вибір алгоритмів обґрунтовано як їхньою загальною ефективністю, так і можливістю імплементації за допомогою сучасних мов програмування, таких як Python.

У **третьому розділі** магістерської роботи **«Апаратно-програмне забезпечення системи»** описано вибір стеку технологій для реалізації модулів скорингу клієнтів та реалізація обох модулів. Програмний комплекс

реалізовано мовою програмування Python, з використанням бібліотеки sklearn для модуля скорингу клієнтів, та resemblyzer для модуля розпізнавання мовця.

Діаграму станів розробленої програми наведено на рис. 1.

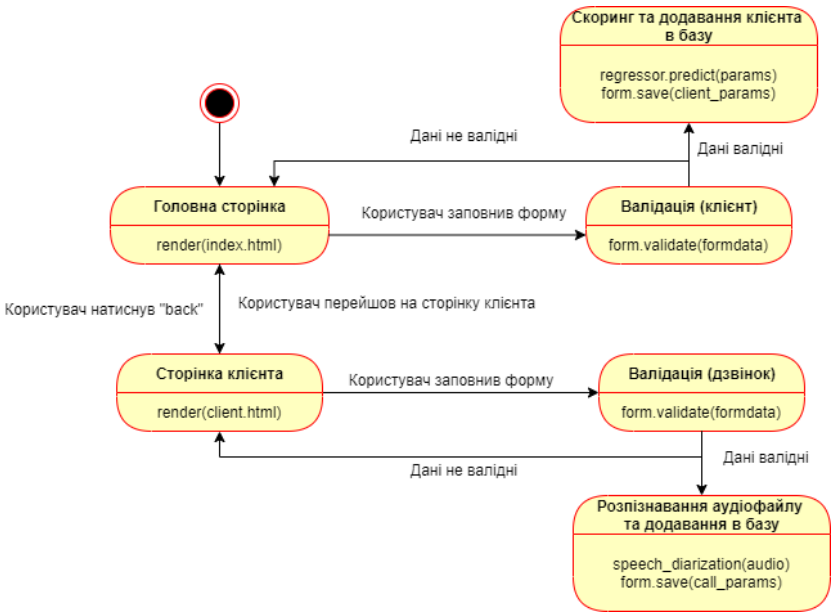

Рисунок 1 – Діаграма станів розробленої програми

В якості алгоритму скорингу, обрано алгоритм Random Forest, з наступними значеннями основних метрик якості:

– R^2 – 94,5%
– MSE – 2,76
– MAE – 1,24

В четвертому розділі магістерської роботи «**Експериментальні дослідження. Тестування та перевірка працездатності апаратно-програмного забезпечення**» описано розроблення веб-інтерфейсу для простого та швидкого доступу до всіх функцій алгоритму скорингу та автоматичного розпізнавання мовця, а також експериментальні дослідження та процес тестування і перевірки працездатності розробленого користувацького інтерфейсу.

На етапі розроблення користувацького інтерфейсу створено діаграму варіантів використання, специфікації основних варіантів використання, та макети сторінок. Реалізацію веб-інтерфейсу за створеними макетами

виконано з використанням фреймворку Django, мови шаблонів Jinja, та HTML.

Проведене функціональне тестування варіантів використання користувацького інтерфейсу методом тестування за класами еквівалентності. Визначено очікуваний формат вхідних та вихідних даних, та розроблено план тестування, за яким можна перевірити працездатність усіх функцій користувацького інтерфейсу.

Проведено тестування алгоритму автоматичного розпізнавання мовця, реалізованого в ході виконання даної роботи. Встановлено, що рівень похибки алгоритму (DER) становить 14,73%.

Додатки містять лістинги кодів розробленого веб-застосунку, а також кодів алгоритму автоматичного розпізнавання мовця та код для навчання скорингової моделі.

У спеціальній частині «Охорона праці та безпека життєдіяльності» розглянуто умови праці у відділі розробки програмного забезпечення ТОВ «Екстравест», та запропоновано деякі заходи з покращення умов праці на підприємстві.

ВИСНОВКИ

В ході виконання магістерської роботи розроблено апаратно-програмне забезпечення для використання у відділі продажів аутсорсингової IT-компанії. Апаратно-програмне забезпечення складається із двох частин: модуля скорингу потенційних клієнтів IT-компанії та алгоритму його роботи, та модуля автоматичного розпізнавання мовця і його алгоритму.

Показано, що найбільш ефективним алгоритмом скорингу клієнтів є Random Forest, оскільки він є менш упередженим завдяки тому, що всі дерева прийняття рішень, які входять до «лісу», тренуються на різних частинах датасету. За такої переваги відмінність точності алгоритму Random Forest від інших є незначною. Встановлено, що середня похибка алгоритму

складає 1,24 бали за результатами навчання на датасеті з даних про 450 реальних клієнтів аутсорсингової IT-компанії, що вперше здійснено для реалізації скорингу клієнтів в IT-сфері.

Встановлено, що для розпізнавання мовця найбільш точним для конкретного набору ознак є алгоритм формування d-векторів за допомогою мел-частотних кепстральних коефіцієнтів, що потрапляють на вхід попередньо навченої LSTM-мережі. Виявлено, що найбільш ефективною є імплементація алгоритму d-векторів за допомогою мови програмування Python з використанням бібліотек sklearn та resemblyzer, а також фреймворку Django для розроблення веб-інтерфейсу. Експериментальне тестування роботи алгоритму для розпізнавання мовця становить 14,73%.

Розроблена програма забезпечує оброблення завантажених аудіофайлів в форматах .mp3 та .wav, та перетворення усного мовлення у текстовий формат із автоматичним розпізнаванням мовців. Також, розроблена програма забезпечує скоринг клієнтів на основі інформації, введеної у веб-форму.

Основні недоліки розробленого комплексу – похибки під час переведення аудіозаписів переговорів у текстовий формат, а також порівняно довготривалий процес додавання даних про нового клієнта. Під час подальшого розвитку проекту, планується покращити якість розпізнавання, використавши альтернативні методі розпізнавання усного мовлення. Також планується скоротити тривалість процесу додавання даних про нового клієнта шляхом інтеграції з платформою Upwork та отримання даних за допомогою API.

Результати роботи можуть бути рекомендовані до застосування в аутсорсингових IT-компаніях, наприклад Extrawest GmbH.

В процесі виконання роботи також виконано спеціальну частину з охорони праці. Проаналізовано умови роботи у відділі продажів IT-компанії. Для покращення робочих умов, запропоновано встановити більш потужні

освітлювальні пристрої, запровадити інший режим роботи, та більш ефективні моделі менеджменту. Вказані заходи можуть призвести до підвищення продуктивності праці співробітників відділу на 8,42 %.

СПИСОК ПУБЛІКАЦІЙ ЗА ТЕМОЮ РОБОТИ

1. Овчар К. В., Дворник О. В. Чорноморський національний університет ім. Петра Могили «Використання методів регресійного аналізу та автоматичного розпізнавання мовця для вдосконалення процесу продажів в ІТ сфері».

АНОТАЦІЯ

Овчар К. В. Скорингова система заключення контрактів в ІТ-сфері. – Кваліфікаційна робота магістра зі спеціальності 123 Комп'ютерна інженерія. – Чорноморський національний університет імені Петра Могили, 2021.

Магістерська робота присвячена розробленню програмного комплексу з модулями автоматичного скорингу потенційних клієнтів аутсорсингових ІТ-компаній та автоматичного розпізнавання мовця для збереження записів переговорів із клієнтами у текстовому форматі. Практичне значення результатів дослідження та розроблення полягає у можливості їх використання в аутсорсингових ІТ-компаніях для оптимізації продажів.

Пояснювальна записка магістерської роботи складається зі вступу, чотирьох розділів, висновків, п'яти додатків та спеціальної частини з охорони праці та безпеки життєдіяльності. У вступі визначається актуальність теми, сформульовані мета, об'єкт, предмет та завдання дослідження та розроблення магістерської роботи. У першому розділі досліджуються наукові публікації, що описують рішення автоматичного скорингу клієнтів та методи розпізнавання мовця. У другому розділі наведено дані про математичну основу розроблених алгоритмів та розглянуто проектування системи. У третьому розділі наведені дані про

практичну реалізацію програмного комплексу. У четвертому розділі описано процес створення веб-інтерфейсу та тестування розробленого програмного комплексу. У висновках наведено аналіз виконаної роботи та отриманих результатів дослідження та розроблення. У додатках А, Б, та В, Г та Д наведено програмний код, що використовувався в проекті. Спеціальна частина присвячена охороні праці на підприємстві, де проводилось тестування комплексу.

В цілому, магістерська робота без додатків містить 71 сторінку, 20 рисунків, 15 таблиць, 31 джерело посилання.

Ключові слова: автоматичний скоринг клієнтів, автоматичне визначення мовця, d-вектори, Random Forest, MFCC, спектральна кластеризація, розпізнавання усної мови, Django.

ABSTRACT

Ovchar K. Scoring system for automated prediction of deal closing in the IT industry. – Master's thesis in specialty 123 Computer Engineering. – Petro Mohyla Black Sea National University, 2021.

The master's thesis is dedicated to the development of a software system, which includes the modules for automated scoring of prospective clients of outsourcing IT-companies and automated speaker diarization for saving the records of negotiations with the clients in the text format. The practical value of the investigation and development results is the possibility to use them at outsourcing IT-companies for sales optimization.

An explanatory note of the master's thesis consists of an introduction, four chapters, conclusions, five annexes, and the special part, which is devoted to the labor protection. The introduction determines the relevance of the topic, formulates the purpose, object, subject and objectives of the research and development of this master's thesis. The first chapter is devoted to investigation of the scientific publications, where existing solutions for automated client scoring

and speaker diarization are described. The second chapter describes the mathematical foundations of algorithms created. The design of the software system is also considered. The third chapter is devoted to the implementation of the software system. The fourth chapter describes web interface development, as well as testing of the software system. The conclusions give an analysis of the work performed and the results of research and development. Annexes A, B, C, D, E contain the program code used in the project. The special part is devoted to the labor protection at the enterprise where the complex was tested.

In general, the master's thesis without annexes contains 71 pages, 20 pictures, 15 tables, 31 reference sources.

Keywords: automated client scoring, speaker diarization, d-vectors, Random Forest, MFCC, spectral clustering, speech recognition, Django.