МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ЧОРНОМОРСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ПЕТРА МОГИЛИ

Арюпін Денис Олексійович

УДК 004.85

ІНФОРМАЦІЙНА СИСТЕМА ПРОГНОЗУВАННЯ ЦІН НА НЕРУХОМІСТЬ НА ОСНОВІ МЕТОДІВ МАШИННОГО НАВЧАННЯ

122 – Комп’ютерні науки

Автореферат
магістерської кваліфікаційної роботи на здобуття освітньої кваліфікації
«Магістр комп’ютерних наук»

Миколаїв – 2021

Магістерська кваліфікаційна робота є рукопис.

Робота виконана в Чорноморському національному університеті імені Петра Могили Міністерства освіти і науки України на кафедрі інтелектуальних інформаційних систем

Науковий керівник: д.т.н., доцент, професор кафедри інтелектуальних інформаційних систем Гожий Олександр Петрович

Рецензент: к.ф.-м.н., доцент кафедри кафедри комп'ютерної інженерії Пузирьов Сергій Володимирович

Захист відбудеться «22» лютого 2021 р. о 9^{30} год. на засіданні екзаменаційної комісії (ауд. 2-403) у Чорноморському національному університеті імені Петра Могили за адресою: 54003, м. Миколаїв, вул. 68-ми Десантників, 10.

З магістерською кваліфікаційною роботою можна ознайомитися в бібліотеці Чорноморського національного університету імені Петра Могили за адресою: 54003, м. Миколаїв, вул. 68-ми Десантників, 10.

Автореферат представлений «18» лютого 2021 р.

Секретар
екзаменаційної комісії,
к.пед.н., доцент Н. М. Болюбаш

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність дослідження визначається тим, що прогнозування невизначеного майбутнього дозволяє знизити ступінь ризику та адекватно керувати бізнес процесами. Правильний прогноз на ринку нерухомості має велике значення для всіх його учасників.

Метою магістерської кваліфікаційної роботи є дослідження та створення моделей прогнозування цін на нерухомість, на основі методів машинного навчання.

Об'єктом процеси прогнозування структурованих даних.

Предметом методи регресійного аналізу та методи прогнозування, засновані на деревах рішень.

Практичне значення даної магістерської кваліфікаційної роботи полягає у можливості застосування моделі для прогнозування цін на нерухомість.

Результати даної магістерської кваліфікаційної роботи було надруковано у тезах XXIII Всеукраїнської науково-методичної конференції «Могилянські читання – 2020» у секції Комп'ютерні науки.

Магістерська кваліфікаційна робота складається із вступу, 3 розділів, висновків, додатків. Загальний обсяг роботи складає 72 сторінки, 23 рисунків, 3 таблиці та 24 посилань на літературні джерела.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У вступі магістерської кваліфікаційної роботи обґрунтовано актуальність обраної теми, сформульовано мету і задачі дослідження, визначено предмет та об’єкт дослідження.

У першому розділі наведено огляд предметної області та аналіз технологій машинного навчання.

У другому розділі здійснено опис усіх технічних засобів та інструментів для розробки прогнозних моделей. Було досліджено методи прогнозування структурованих даних.

Основною бібліотекою для моделі виступає Scikit-learn – один з найбільш широко використовуваних пакетів Python для Data Science і Machine Learning.

Scikit-learn спеціалізується на алгоритмах машинного навчання для вирішення задач навчання з учителем: класифікації (прогноз ознаки, множина допустимих значень якого обмежена) і регресії (прогноз ознаки з дійсними значеннями), а також для задач навчання без учителя: кластеризації (розбиття даних по класах , які модель визначить сама), зниження розмірності (подання даних в просторі меншої розмірності з мінімальними втратами корисної інформації) і детектування аномалій.

Бібліотека реалізує такі основні методи:

- лінійні: моделі, завдання яких побудувати розділяючу (для класифікації) або аппроксимуючу (для регресії) гіперплощину;
- метричні: моделі, які обчислюють відстань по одній з метрик між об'єктами вибірки, і приймають рішення в залежності від цієї відстані (K найближчих сусідів);
- дерева рішень: навчання моделей, що базуються на множині умов, оптимально обраних для вирішення завдання;
- ансамблеві методи: методи, засновані на деревах рішень, які комбінують можливості безлічі дерев, і таким чином підвищують їх якість роботи, а також дозволяють проводити відбір ознак (бустінг, беггінг, випадковий ліс, мажоритарне голосування);

- нейронні мережі: комплексний нелінійний метод для задач регресії і класифікації;
- SVM: нелінійний метод, який навчається визначати межі прийняття рішень;
- PCA: лінійний метод зниження розмірності і відбору ознак;
- t-SNE: нелінійний метод зниження розмірності;
- K-середніх: найпоширеніший метод для кластеризації, що отримує на вхід число кластерів, за якими повинні бути розподілені дані;
- крос-валідація: метод, при якому для навчання використовується весь датасета (на відміну від розбиття на вибірки train / test), проте навчання відбувається багаторазово, і в якості валідаційної вибірки на кожному кроці виступають різні частини датасета. Підсумковий результат уявляють собою усереднення отриманих результатів;
- grid search: метод для знаходження оптимальних гіперпараметров моделі шляхом побудови сітки з значень гіперпараметрів і послідовного навчання моделей з усіма можливими комбінаціями гіперпараметров з сітки.

Розглянуто методи лінійної регресії та випадкового лісу.

У математичній статистиці лінійна регресія є метод апроксимації залежностей між вхідними та вихідними змінними на основі лінійної моделі. Є частиною більш широкої статистичної методики, званої регресійний аналізом.

У регресійному аналізі вхідні (незалежні) змінні називаються також предикторними змінними або регресорами, а залежні змінні - критеріальними.

Якщо розглядається залежність між однією вхідною і однією вихідною змінними, то має місце проста лінійна регресія. Для цього визначається рівняння регресії і будується лінія регресії. Якщо ж шукається залежність між декількома вхідними і однією вихідною змінними, то має місце множинна лінійна регресія. Відповідне рівняння має вигляд:

$$Y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n \qquad (2.1)$$

де n – кількість змінних.

Відмінність між простою і множинною лінійною регресією полягає в тому, що замість лінії регресії в ній використовується гіперплощина.

Перевага множинної лінійної регресії в порівнянні з простою полягає в тому, що використання в моделі декількох вхідних змінних дозволяє збільшити частку пояснений дисперсії вихідної змінної, і таким чином поліпшити відповідність моделі даних. Тобто при додаванні в модель кожної нової змінної коефіцієнт детермінації зростає.

Метод випадкового лісу - один з найбільш універсальних алгоритмів машинного навчання. Універсальність полягає, по-перше, в тому, що він добре показує себе в багатьох задачах, по-друге, в тому, що є випадковий ліс для вирішення завдань класифікації, регресії, кластеризації, пошуку аномалій, селекції ознак і т.д.

Метод швидко отримав визнання як в статистичному співтоваристві, так і в середовищі дослідників, що використовують методи розпізнавання образів в своїй роботі і в даний час є одним з найбільш популярних методів класифікації і непараметричної регресії. Причиною цього стала не тільки висока точність класифікації, що забезпечується методом, а й інші його гідності. Саме:

- метод гарантує захист від перенавчання навіть в разі, коли кількість ознак значно перевищує кількість спостережень. Це властивість виділяє метод "випадковий ліс" серед безлічі інших методів класифікації і є надзвичайно цінним для вирішення багатьох прикладних задач;
- для побудови випадкового лісу по навчальній вибірці потрібно завдання всьогодвох параметрів, які вимагають мінімальної налаштування;
- випадкові ліси можуть використовуватися не тільки для задач класифікації і регресії, а й для задач виявлення найбільш інформативних ознак, кластеризації, виділення аномальних спостережень і визначення прототипів класів;
- навчальна вибірка для побудови випадкового лісу може містити ознаки, виміряні в різних шкалах: числовій, порядковій і номінальній, що неприпустимо для багатьох інших класифікаторів;

- метод допускає легку паралелізацію (програмну реалізацію, придатну для паралельних обчислень), що має велике значення при великих обсягах навчальної вибірки.

Випадковий ліс - це множина вирішальних дерев. У задачі регресії їх відповіді усереднюються, в завдані класифікації приймається рішення голосуванням за більшістю. Всі дерева будуються незалежно.

У третьому розділі наводиться опис розробленої прогнозної моделі.

У спеціальній частині магістерської кваліфікаційної роботи з «Охорони праці та безпеки життєдіяльності» розглянуто мікрокліматичні умови праці на робочих місцях на предмет виробничого освітлення та дотримання вимог експлуатації ПК, а також впроваджені заходи щодо запобігання захворюванню на коронавірус. Аналіз умов праці в розглянутому робочому приміщенні показав, що умови праці з ПЕОМ відповідають вимогам . Приміщення, в якому розташовано робоче місце має достатні площу та об’єм для роботи однієї людини. Завдяки використанню сучасного обладнання та підбору оптимальної продуктивності комп’ютера відповідно до роботи, що виконується, рівень шуму комп’ютера не перевищує нормативні значення.

Аналіз параметрів мікроклімату показав, що вони не в повній мірі відповідають вимогам нормативних документів. В зимовий період вологість повітря знаходиться на межі допустимих значень, а в теплу пору року температура повітря на робочому місці може перевищувати норму. Для приведення мікроклімату до відповідності нормам необхідне застосування зволожувача повітря у зимовий період, та кондиціонера у теплу пору року.

Ергономіка робочого місця і режим зорової роботи задовольняють вимогам і сприяють зниженню втоми. Для збереження здоров'я робітника, запобігання професійним захворювання і підтримки працездатності необхідно суворо дотримуватись вимог до режимів праці і відпочинку при роботі з ВДТ ЕОМ і ПЕОМ, що викладені у пункті 5 [1].

Аналіз стану щодо епідемії коронавірусу показав, що впроваджено усі необхідні заходи для безпеки та запобіганню розповсюдження коронавірусної інфекції.

У методичній частині розроблено практичну роботу на тему «Дерева рішень».

ЗАГАЛЬНІ ВИСНОВКИ

У даній магістерські кваліфікаційній роботі досліджено та створено модель прогнозування цін на нерухомість. У першому розділі наведено огляд предметної області та аналіз технологій машинного навчання. У другому розділі здійснено опис усіх технічних засобів та інструментів для розробки прогнозних моделей. Було досліджено методи прогнозування структурованих даних такі як лінійна регресія і випадковий ліс. Також було досліджено методи та критерії щодо оцінки якості моделі та прогнозу. Такі критерії як коефіцієнт детермінації, залишкова сума квадратів, критерій Дарбіна – Уотсона, середньоквадратична похибка, середня абсолютна и середня абсолюна відсоткова похибки. В результаті цього аналізу, у третьому розділі розроблено прогнозну модель прогнозування цін на нерухомість.

У методичній частині магістерської роботи розроблено практичну роботу на тему «Дерева рішень».

У спеціальній частині магістерської роботи з «Охорони праці та безпеки в надзвичайних ситуаціях» здійснено аналіз умов праці на робочому місці. Виконано перевірочний розрахунок природного освітлення та розраховано загальне рівномірне освітлення люмінесцентними лампами в розглянутому приміщенні. Розглянуто запроваджені дії для запобігання захворюванню на коронавірус.

АНОТАЦІЯ

Арюпін Денис Олексійович. Інформаційна система прогнозування цін на нерухомість на основі методів машинного навчання. – На правах рукопису.

Магістерська кваліфікаційна робота на здобуття освітньої кваліфікації «Магістр комп'ютерних наук». – Чорноморський національний університет імені Петра Могили, Миколаїв, 2021.

Визначення та прогнозування цін на нерухомість має велике значення для аналізу динаміки цін на ринку нерухомості та для більш точного визначення та регулювання запитів на різні види нерухомості. Однак ручний аналіз таких великих об'ємів даних буде занадто складним та потребуватиме величезних обсягів часу. Задля цього можна використовувати технології машиного навчання.

Мета роботи - дослідження та створення моделей прогнозування цін на нерухомість, на основі методів машинного навчання.

Фахова частина містить вступ, три розділи, висновки та додатки до кваліфікаційної роботи.

У першому розділі було проведено аналіз прдеметної області.

У другому розділі було обрано технічні засоби та інструменти для розробки прогнозних моделей. Було досліджено методи прогнозування структурованих даних.

У третьому розділі було виконано порівняльний аналіз прогнозної моделі та критеріїв якості прогнозу та розроблено програмне забезпечення для реалізацій прогнозних моделей.

Кваліфікаційна робота містить: сторінок – 72, рисунків – 23, таблиць – 3 додатків – 1, посилань – 24.

Ключові слова: машинне навчання, модель, прогнозування, оцінка якості моделі.

ABSTRACT

Aryupin Denis Alekseevich. Information system for forecasting real estate prices based on machine learning methods. – Have the rights of the manuscript

Diploma work for receiving educational qualification “Master of Computer Science”. - Petro Mohyla Black Sea National University, Mykolayiv, 2021

Determining and forecasting real estate prices is of great importance for analyzing the dynamics of prices in the real estate market and for more accurate definition and regulation of requests for different types of real estate. However, manually analyzing such large amounts of data will be too complex and time consuming. To do this, you can use machine learning technology.

The purpose of the work - research and creation of models for forecasting real estate prices, based on machine learning methods.

The professional part contains an introduction, three sections, conclusions and appendices to the qualification work.

In the first section, an analysis of the subject area was conducted.

In the second section, technical means and tools for the development of forecast models were selected. Methods for predicting structured data were investigated.

In the third section, a comparative analysis of the forecast model and forecast quality criteria was performed and software for the implementation of forecast models was developed.

Qualification work contains: pages - 72, figures - 23, tables – 3, appendices – 1, links - 24.

Keywords: machine learning, model, forecasting, model quality assessment