НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ

«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ імені ІГОРЯ СІКОРСЬКОГО»

Кафедра технічної кібернетики

ПОЯСНЮВАЛЬНА ЗАПИСКА ДО КУРСОВОЇ РОБОТИ

з дисципліни «Аналіз даних в інформаційно-управляючих системах»

на тему

«Прогноз цін на нерухомість у Бостоні»

«ВИКОНАВ»

студент групи ІТ-92

Щур Антон Сергійович

ПІБ

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Підпис                    Дата

«ПРИЙНЯВ» з оцінкою

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

доц. Ліхоузова Т.А. / ас. Вітюк А.Є.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Підпис                    Дата

Зміст

1. Загальна характеристика предметної області та постановка задачі…….3

* 1. Огляд предметної області ...………………………………………...3
  2. Огляд доступних джерел даних ……………………………………3

1.3 Постановка задачі …………………………………………………...8

2. Вибір моделей ……………………………………………………………..9

* 1. Визначення моделей, що можуть бути використані ……………..9
  2. Вибір ознак, що будуть використані для аналізу ………………..9
  3. Підготовка даних для навчання та верифікації моделей ………..12
  4. Формування моделей. ……………………………………………..13
  5. Верифікація моделей ………………………………………………17
  6. Висновки щодо якості побудованих моделей …………………...19

1. Результати аналізу ……………………………………………………….20

1. Загальна характеристика предметної області та постановка задачі
   1. Огляд предметної області

Прогнозування ціни на будь-який товар чи послугу є досить необхідною річчю в нашому житті. В моєму випадку це прогнозування ціноутворення на нерухомість. Точний прогноз допоможе будівельній компанії або людині, що захоче продати своє житло, зробити це максимально вигідно без великих втрат по часу або в капіталі. Зазвичай люди, вибираючи собі житло, дивляться на багато факторів таких як площа, кількість кімнат, дехто вибирає ще поверх квартири. Сьогодні на ціну житла можуть впливати багато факторів, що заховані від неозброєного ока. Тому, від даного дослідження я хочу навчитись більш менш точно прогнозувати ціну на житло, а також виявити фактори які неявно впливають на ціноутворення нерухомості.

1.2 Огляд доступних джерел даних

Шукаючи дані в інтернет ресурсах я наткнувся на досить цікавий сайт Kaggle.com. Як виявилося це платформа для змагань. На даному ресурсі я знайшов необхідні для мене дані, тобто датасет, що має назву Boston House Prices. Даний датасет був зібраний у 2004 році. Також на даному ресурсі був опис даного датасету. Загалом в даних існує 14 стовбців з варіантами та 506 рядків з даними.

Перейдемо до опису стовбців датасету:

CRIM – рівень злочинності на душу населення.

ZN – частка житлової землі у %, якщо площа займає більше ніж 25 тис квадратних футів.

INDUS – частка акрів, що підлягають нероздрібній торгівлі.

CHAS – 1, якщо ділянка межує з річкою Чарльз, 0 – не межує.

NOX – концентрація оксиду азоту.

RM – середня кількість кімнат на житло.

AGE – частка зайнятих будинків у відсотках, побудованих до 1940 року.

DIS – зважені відстані до центрів зайнятості Бостона.

RAD – індекс доступності магістралей.

TAX – повна ставка податку на майно на кожні 10 тис. доларів США.

PTRATIO – відношенння учня до вчителя за ройоном міста.

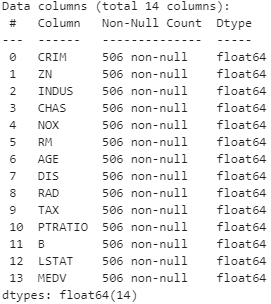
B – 1000(Bk-0.63)^2, де BK – частка чороношкірих жителів району.

LSTAT – % населення з малим рівнем життя.

MEDV – середня ціна на будинок в 1000 доларів США.

В даному датасеті наведено досить цікаві варіанти. Окрім середньої кількості кімнат на житло, я виділив для себе декілька цікавих варіант: LSTAT, NOX, RAD, CRIM. Оскільки, при виборі житла, було б досить непогано звернути увагу на викиди азоту, доступність магістралей, на скільки район є кримінальним та наскільки багаті люди проживають в ньому. Перейдемо до описової статистики нашого датасету.

Переглянимо основну інформацію про наші дані (рис. 1.2.1)

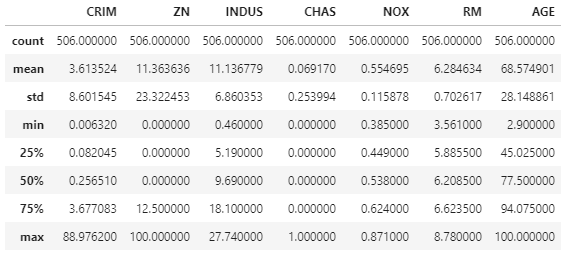


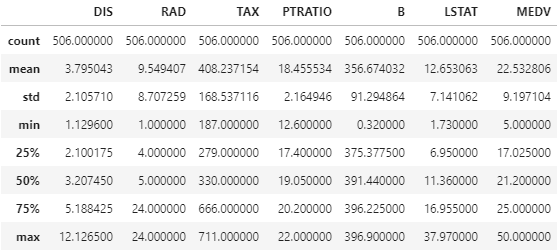
**Рис 1.2.1 Перегляд даних по датасету**

З даного малюнку можемо побачити, що у нас всі стовбці мають тип float, а також немає пустих значень. Тому можемо переходити до описової статистики датасету.

Описова статистика для даних:

На рисунку 1.2.2 показано середнє, СКВ, мінімальне та максимальне значення а також квартилі для кожного стобвця датасету.

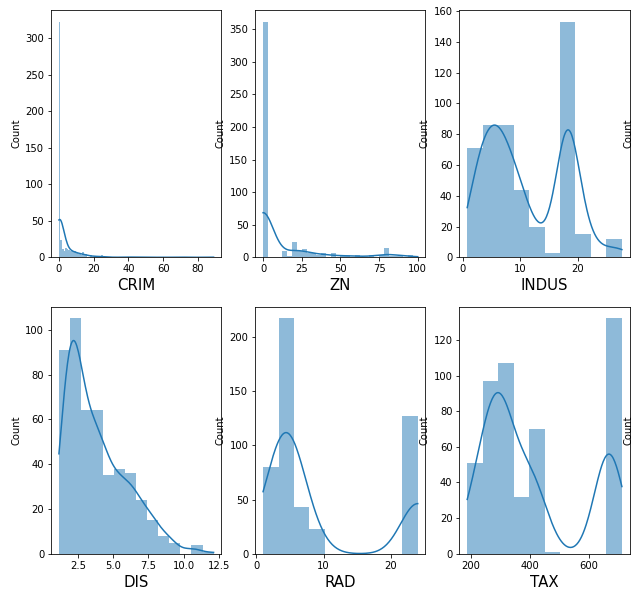


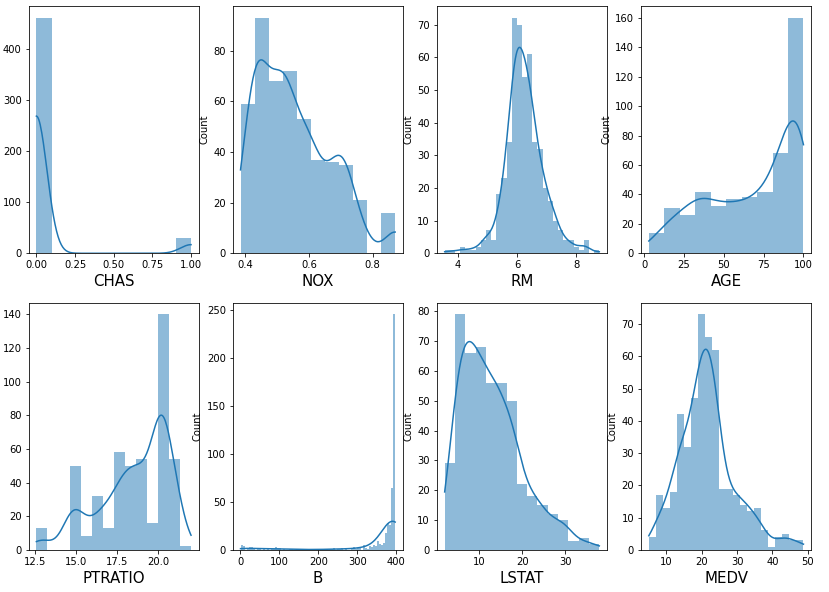


**Рис 1.2.2 Описова статистика датасету**

З даного малюнку видно, що майже всі стобвці мають викиди. Для наглядності побудуємо графіки Histogram та BoxPlot.

На рисунку 1.2.3 зображені гістограми частот для кожного стовбця наших даних.

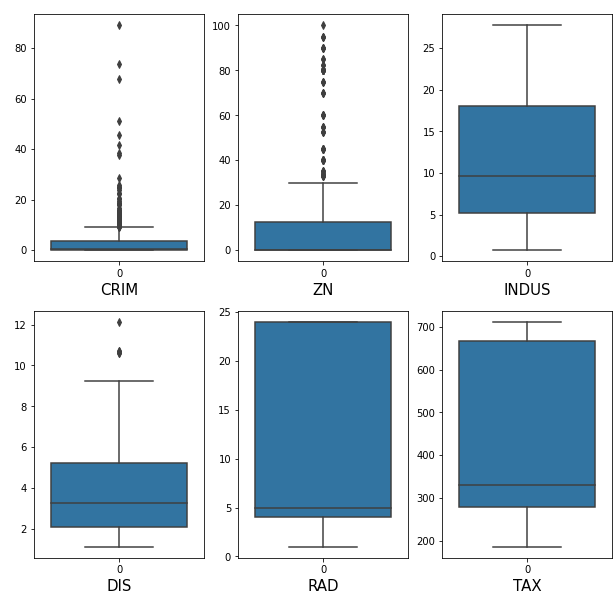


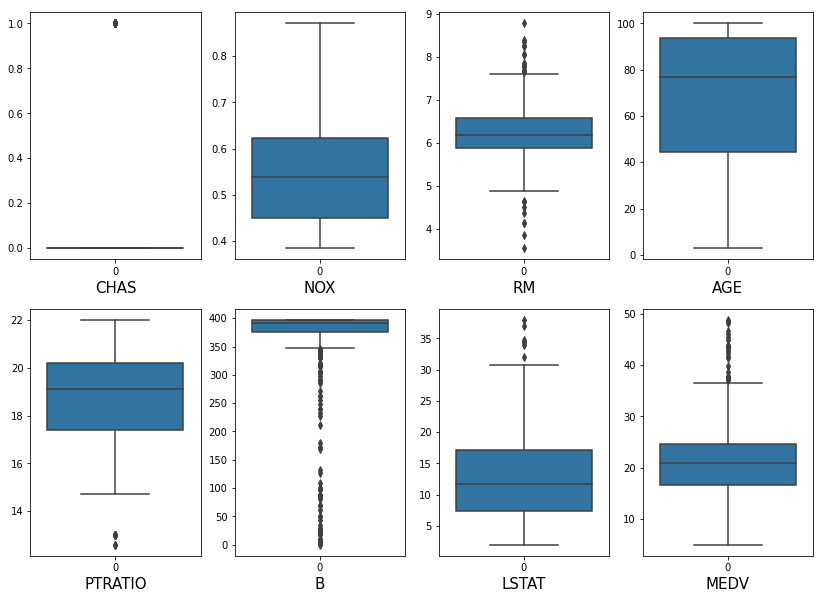


**Рис 1.2.3 Гістограми частот для даних**

З даного малюнку видно, що змінні CHAS, B, CRIM, CHAS – мають досить багато викидів. Також видно, що MEDV, RM – мають вигляд близький до нормального розподілу. Видно, що є і одномодальні розподіли з лівим чи правим хвостом – це NOX, AGE, DIS.

На рисунку 1.2.4 зображені діаграми BoxPlot для кожного стовбчика з даними:





**Рис 1.2.4 Діаграми Boxplot для кожного стовбця датасету**

З даних діаграм я побачив наглядно, що CRIM, ZN, RM, B мають досить багато викидів. На далі особливо буду звертати увагу на дані стовбчики.

1.3 Постановка задачі

Метою мого дослідження є навчитись прогнозувати ціну на нерухомість за вхідними даними, що відповідають структурі обраного мною датасету. Тобто, я хочу прогнозувати ціну, опираючись на такі фактори: zn, indus, chas, nox, rm, age, dis, rad, tax, ptratio, lstat та b.

Питання, на які має відповісти моє дослідження:

1. Який фактор або фактори найбільше впливають на ціну житла?
2. Який фактор або фактори найменше впливають на ціну житла?
3. Чи буде краще прогнозування ціни, якщо використовувати більш складну модель?
4. Вибір моделей
   1. Визначення моделей, що можуть бути використані

Мої дані містять 13 варіант (features) та одну цільову змінну (target). Тому я буду використовувати метод навчання з вчителем. Цей вид навчання використовується тоді, коли ми знаємо вхідні та вихідні дані і нам потрібно побудувати функцію, яка буде відображати значення з простору предикторів у простір відгуку.

Моє дослідження пов’язане з розв’язком задачі регресійного аналізу. Отже, я буду використовувати регресійні моделі. В даному досліджені я планую взяти лінійну та поліноміальну регресію. Дерева рішень та випадковий ліс у варіації регресора теж підійдуть для нашого дослідження. Також хочу взяти алгоритм К найближчих сусідів у варіації регресора.

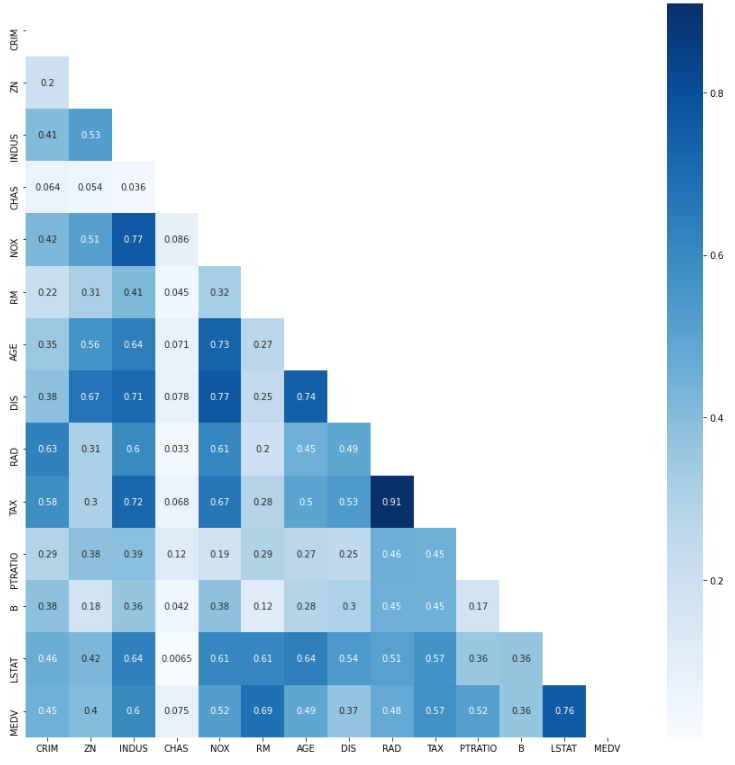
2.2 Вибір ознак, що будуть використані для аналізу

За попереднім аналізом предикторів, я побачив що у мене є категоріальна змінна, це CHAS. Вона показує, чи межує ділянка з річкою Чарльз. Оскільки дана річка проходить не через всі ділянки, і охоплює лише малу частку ділянок вона може нам не підійти. Перевіримо її при подальшому аналізі.

Для відбору факторів при регресійному аналізі може дуже допомогти теплова кореляційна карта.

Побудуємо кореляціїну матрицю:

На рисунку 2.2.1 зображено теплову кореляційну карту для моїх даних.

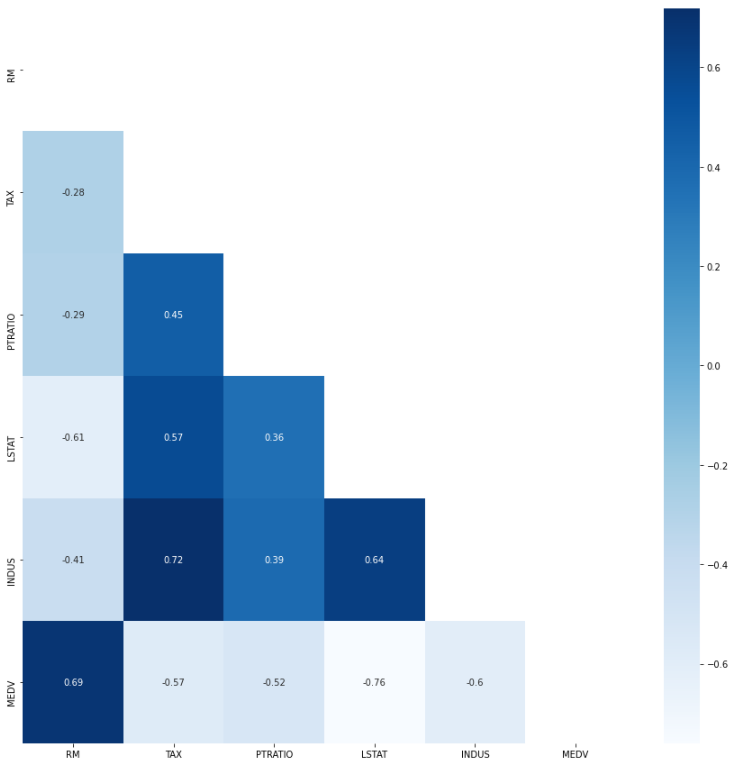


**Рис 2.2.1 Теплова кореляційна карта для усіх предикторів**

Я вивів для зручності карту абсолютних значеннь, оскільки, на даному етапі, мені не важливо позитивна чи негативна кореляція між предикторами, зараз мені важливо зрозуміти, які фактори найбільше впливають на цільову змінну. Отже, переглянувши дану матрицю, можна сказати, що на MEDV (цільова величина) найбільше впливають такі фактори: LSTAT, TAX, PTRATIO, RAD, RM, INDUS. Також, побачив сильну кореляцію між RAD та TAX. Для мене не дуже зрозуміла залежність податку та дальністю розташування магістралей. Така сильна кореляція призводить до мультиколінеарності. Тому для вирішення даного питання, потрібно обрати лише один фактор з цих двох. Побачивши, що TAX потужніше впливає на ціну, я залишу його. Така сама ситуація з NOX та INDUS, вони теж досить потужно корелюють між собою, це логічно, оскільки, на мою думку, INDUS позначає промисловість, хоч це не вказано в описі датасету. Промисловість є забруднювачем повітря. Тому залишу предиктор INDUS. Отже, після відбору залишились такі фактори: LSTAT, TAX, PTRATIO, RM, INDUS.

Подивимось знову на кореляційну матрицю:

На рисунку 2.2.2 зображена нова кореляційна карта, після відбору предикторів.



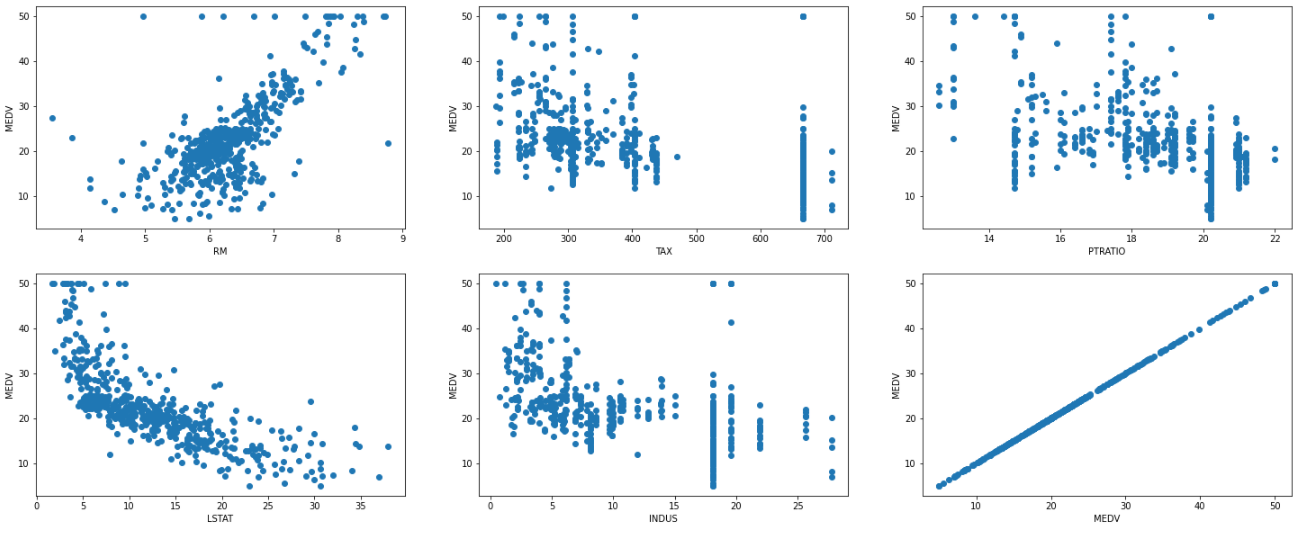
**Рис 2.2.2 Нова теплова кореляційна карта**

З даної карти можна додатково побачити, що тільки кількість квартир позитивно впливає на вартість житла. Високі податки, низький рівень життя, заводи та фабрики, очевидно, мають негативну кореляцію і знижують вартість житла. Також відношення кількості учнів до вчителів впливає негативно на ціну, на мою думку це логічно, оскільки постає проблема навчання дітей і недостачі вчителів. Якщо вчителів мало, підвищується їх навантаженість та понижується концентрація уваги на одному учню, закономірно падає рівень шкільної освіти.

2.3 Підготовка даних для навчання та верифікації моделей

При виконанні описової статистики я помітив, що пропущених значень немає. На рахунок викидів я помітив, що в деяких стовбчиках вони є і їх дуже багато, але ми відкинули ці стовбчики на минулому етапі відбору предикторів. Хоча я помітив деяку закономірність у викидах. Для цього побудуємо діаграму розсіювання для предикторів, що залишились

На рисунку 2.3.1 наведено діаграму розсіювання ціни і кожного предиктора.



**Рис 2.3.1 Діаграма розсіювання для ключових предикторів**

З даного рисунку я побачив, що при ціні в 50 тис. $ США знаходиться багато викидів. Наприклад візьмемо середню кількість кімнат. Можна побачити, що інколи просять доволі велику суму за 5-ти кімнатне житло, хоча в середньому сума сягає приблизно 20 тис. $ США. Так само з іншими предикторами, можна бачити, що ціна обмежена 50 тис. Отже, я прийняв рішення про видалення усіх записів у яких ціна дорівнює 50 тис.

Я вирішив нормалізувати всі мої дані, окрім ціни, зробивши це шляхом віднімання середнього і ділення на СКВ. Отже, я отримую мої дані в Z-шкалі. Тобто маю середнє 0 і СКВ 1. На мою думку це полегшить навчання для моїх моделей та зробить їх точнішими.

Перед навчанням я вирішив розбити мою вибірку на навчальну та тестову. Для тестової вибірки я відвів 20% всього об’єму вибірки.

Також, я буду використовувати метод крос-валідації. Я вибрав 10 k-Fold для моїх даних, тобто я на кожному етапі відкидую 10 % вибірки на тест і 90 % на навчання. Так я проходжу по всьому тренувальному масиву.

Критерії, які я буду використовувати – це MAE та R^2. Я взяв ці метрики для того, щоб бачити точність моделі а також на скільки мої дані далекі від правильних. Скоригований R^2 тут не потрібний, оскільки ми використовуємо моделі з однаковою кількістю вхідних предикторів. Середня абсолютна помилка покаже на скільки ми всередньому відхиляємось від правильного значення.

2.4 Формування моделей. Вибір оптимального класу складності моделей

Для створення та навчання моделей я використовую мову python та пакет для машиного навчання scikit-learn.

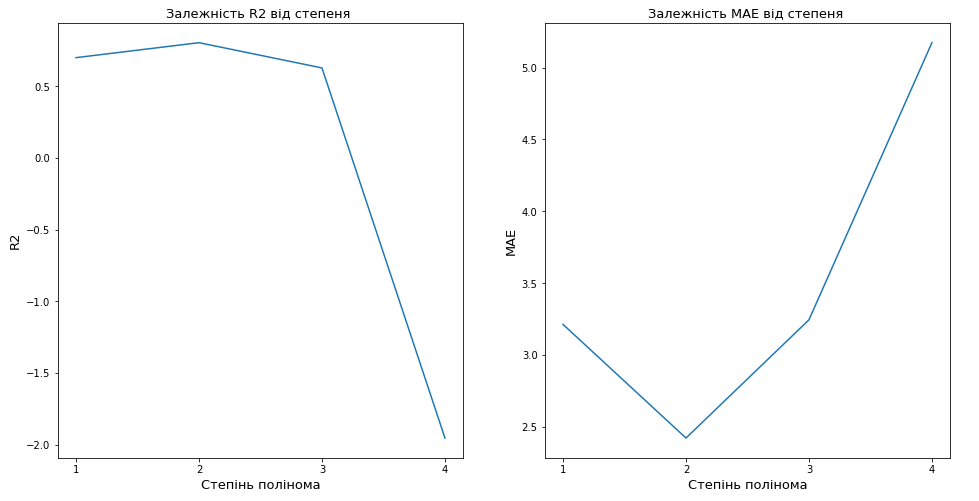
**Лінійна регресія.**

Для лінійної регресії не передбачено ніяких покращень, окрім вибору оптимізатора та регуляризацій L1 та L2. Оскільки я взяв дану модель як найпростішу, я вирішив залишити її в стандартному стані.

**Поліноміальна регресія.**

Даний вид регресії є вже більш гнучким і складнішим. Тут нам необхідно обрати степінь полінома, який буде використано при цій регресії. Тому нам потрібно подивитись, який степінь підійде найкраще. При виборі степеня я користуюсь метрикою MAE та R2. Нам необіхдно підібрати степінь так, щоб не перенавчити нашу модель, оскільки вона просто пристосується до наших навчальних даних і на тестових буде прогнозувати набагато гірше.

Провівши досліди для визначення необхідного оптимального степеня я отримав такі дані, вони зображені на рисунку 2.4.1.



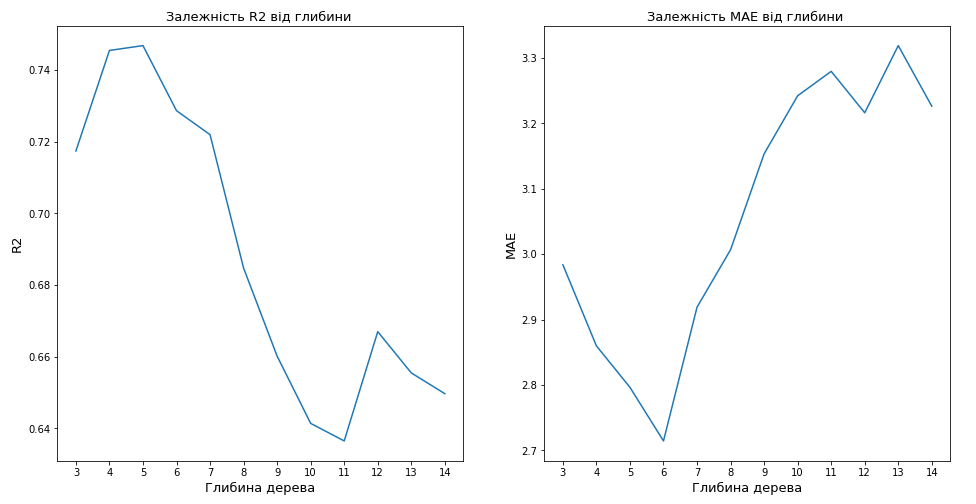
**Рис 2.4.1 Графіки залежності метрики від степеня**

З даного досліду видно, що оптимальний степінь полінома є 2.

Щодо інших налшатуваннь вони теж залишились стандартними, оскільки вже з бібліотеки налаштовані досить непогано.

**Дерево регресор**

Для дерева нам необхідно вибрати максимальну глибину. Я провів такі ж самі дослідження як і для поліномільної регресії і отримав такі дані (рис. 2.4.2):

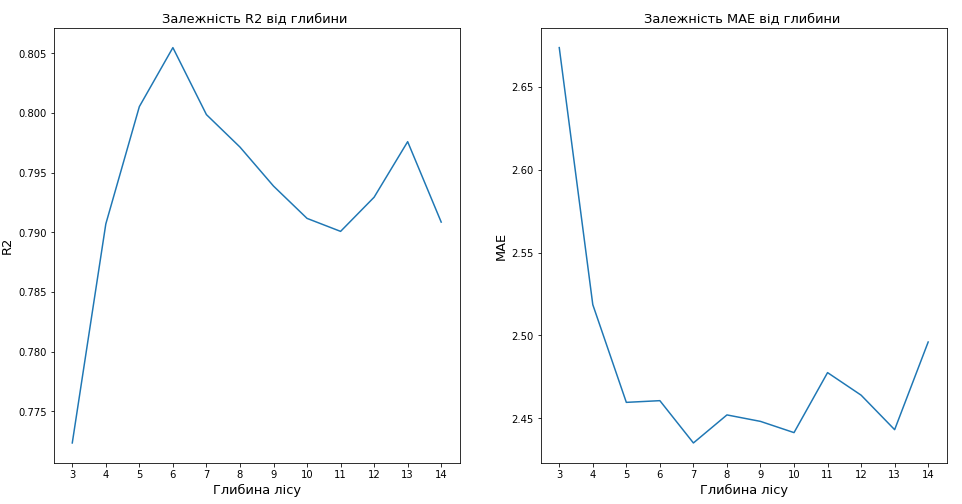


**Рис 2.4.2 Графіки залежності метрик від глибини**

З даного рисунку видно, що нам треба обрати компроміс. Зроблю свій вибір в сторону МАЕ, обираю глибину дерева рівну 6.

**Випадковий ліс**

Для випадкового лісу я теж провів дослідження залежності обраних мною метрик від глибини даного лісу і отримав такі показники (рис. 2.4.3):



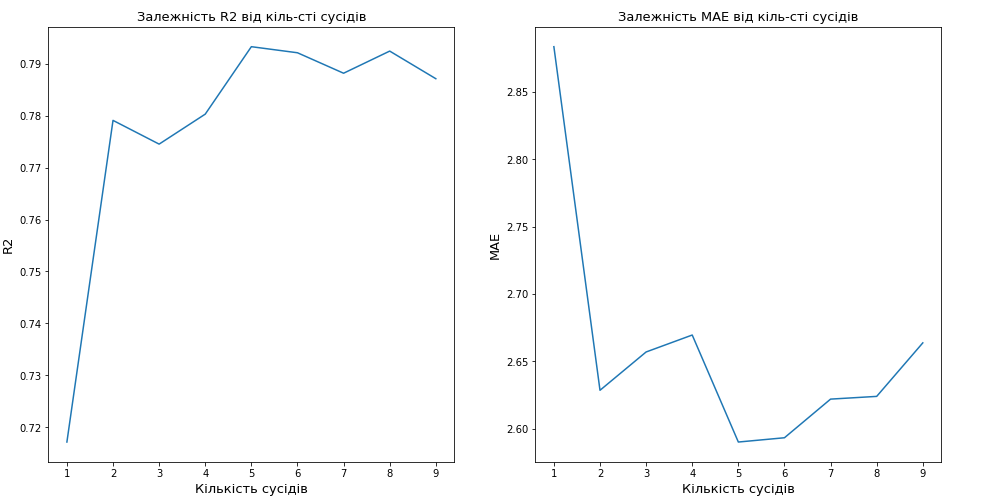
**Рис 2.4.3 Графіки залежності метрик від глибини лісу**

З даного рисунку видно, що нам треба обрати компроміс. Зроблю свій вибір в сторону МАЕ, обираю глибину лісу рівну 7.

**К найближчих сусідів**

Для даного алгоритму нам теж потрібно імперично знайти оптимальну кількість сусідів. Проведемо схожі дослідження для даного алгоритму.

Результати досліду зображені на рисунку 2.4.4



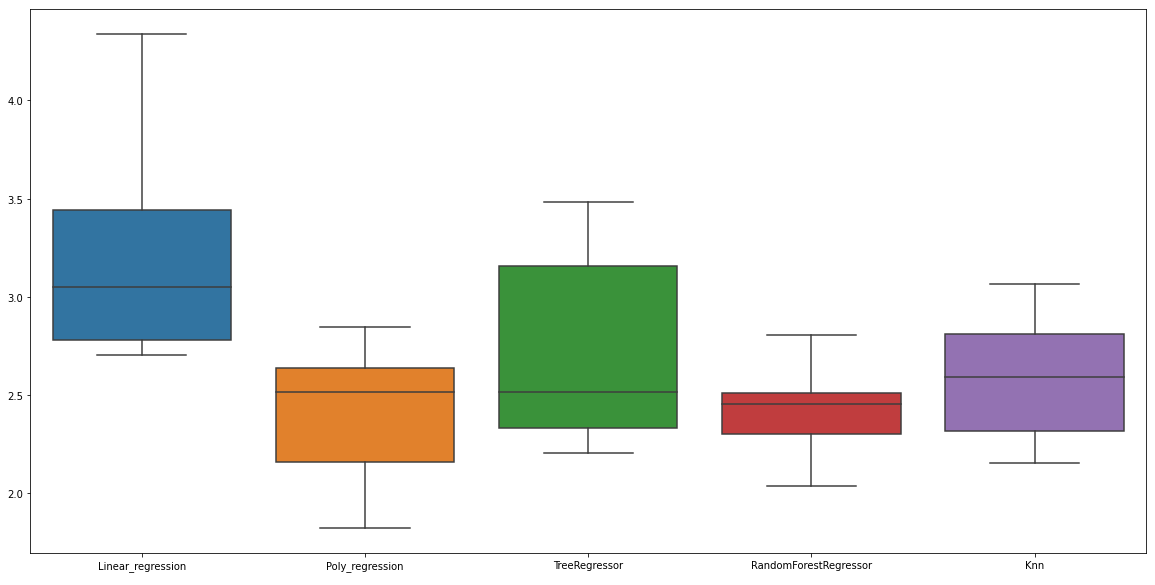
**Рис 2.4.4 Графіки залежності метрик від кількості сусідів**

Вирішив обрати число сусідів рівне 5, оскільки і абсолютна похибка є мінімальною і точність доволі висока.

2.5 Верифікація моделей

Проводивши навчання моделей на навчальних даних за допомогою метода крос-валідації я фіксував значення для обраних мною метрик для кожної моделі. Подивимось на них.

На рисунку 2.5.1 показані графіки результатів для кожної моделі



**Рис 2.5.1 Графік порівняння моделей на навчальних даних з використанням методу cross-validation.**

З даного графіку видно, що на тестових даних найбільш точними є випадковий ліс та поліноміальна регресія. Подивимось на тестових даних, яка модель буде найкращою (табл 2.5.1).

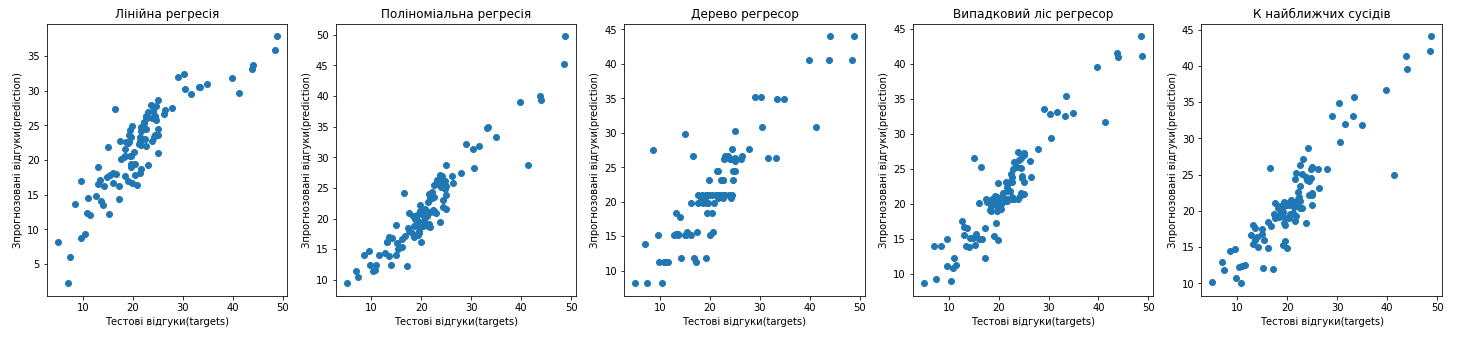
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Модель | Лінійна | Поліноміальна | Дерево | Випадковий  ліс | K найближчих  сусідів |
| MAE | 3.06 | 2.06 | 2.7 | 2.19 | 2.42 |
| R^2 | 0.62 | 0.86 | 0.7 | 0.82 | 0.77 |

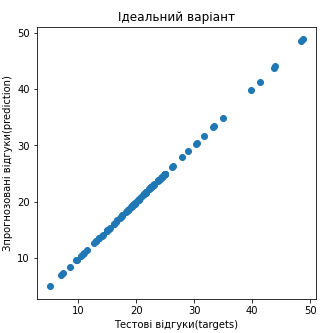
**Таблиця 2.5.1 Порівняння моделей**

З даної таблиці видно, що найкращою з моделей є поліноміальна регресія з степенем полінома рівним 2.

Для наглядності побудуємо діаграми розсіювання, де по осі Х розташуємо тестові відгуки, а по осі У розташуємо зпрогнозовані відгуки.

Дані представлено на рисунку 2.5.2





**Рис 2.5.2 Порівняння різних моделей**

На мою думку, найбільш вдалою з усіх моделей є поліноміальна регресія.

* 1. Висновки щодо якості побудованих моделей.

В даному досліджені я побудував багато різних моделей та навчив їх на навчальних даних з використанням методу крос-валідації, що дозволив мені виділити найбільш вдалі моделі. В лідерах я отримав поліноміальну регресію з степенем полінома рівним 2, також я б виділив і випадковий ліс з глибиною рівною 7. Дані моделі показують доволі схожі результати, тому можна використовувати будь-яку з них, але перевагу я б надав поліноміальній регресії.

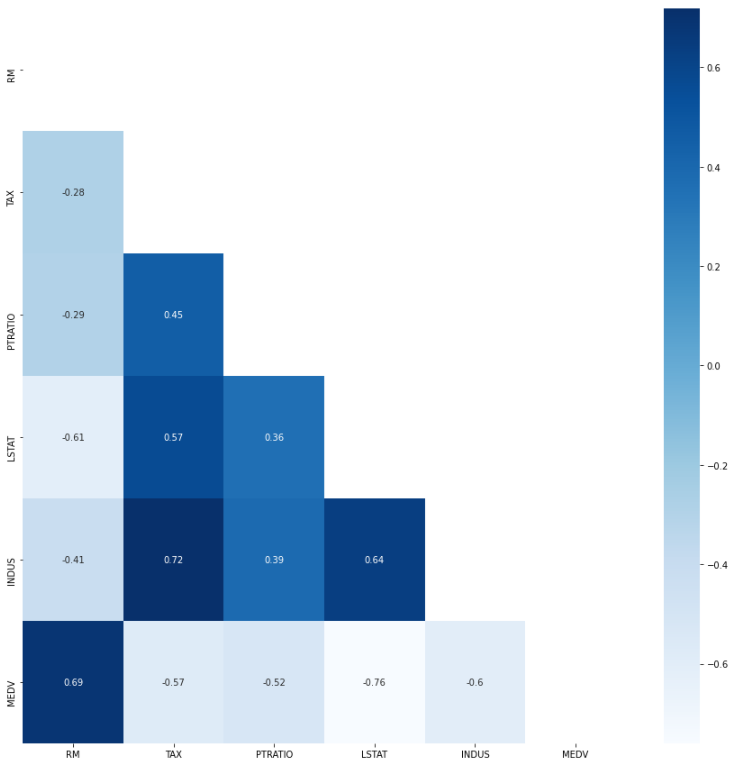
Щодо легкості інтерпритації та покращення моделі при збільшені вибірки, перевагу я б віддав теж поліноміальній регресії, оскільки це більш проста модель і вона показує чудові результати. На мою думку, при збільшені вибірки та її якості, дана модель зможе набагато точніше прогнозувати ціну, хоча це стосується всіх моделей. Вважаючи вищеописані факти, перевагу віддам саме поліноміальній регресії.

1. Результати аналізу

Відповіді на поставлені питання:

1 - Який фактор або фактори найбільше впливають на ціну житла?

Для відповіді на це питання потрібно подивитись на кореляційну теплову карту, яку я використовував на початку аналізу. Для наглядності намалюю її ще раз.

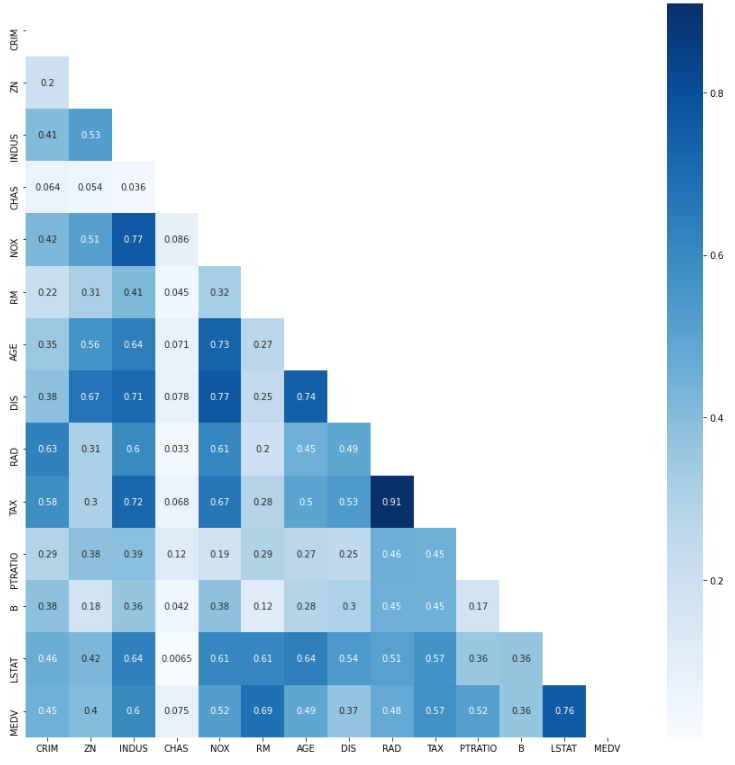


Можна побачити, що максимально негативно впливає на ціну (тобто знижує вартість житла) досить багато факторів, а саме: податки, співвідношення учнів до вчителів, рівень людей з низьким соціальним статусом та наявність промисловості неподалік будику. На мою думку, це досить реально і піддається здоровому глузду. Ніхто не хоте платити багато податків, рівень освіти при відсутності вчителів та при великому навантажені теж падає, промисловість біля житла створює великий шум та забруднєю повітря, а велика кількість бідних людей створює погану атмосферу настрою.

Щодо максимального позитивного впливу на ціноутворення, можна побачити що кількість кімнат досить непогано впливає на ціну житла. Думаю, що це теж дуже реально.

2 - Який фактор або фактори найменше впливають на ціну житла?

Для відповіді на це питання подивимось на повну кореляційну теплову карту



Можна виділити фактор CHAS, наприклад річка Чарльз майже не впливає на ціну. Також частка темношкірих та кримінальність району не дуже сильно впливають на вартість житла. На рахунок кримінальності я маю лише здогадку, оскільки датасет зібраний у Бостоні (США), а в США поліція діє досить добре і швидко і оберігає людей. Також частка чорношкірих не має бути принятою до уваги у толерантних штатах Америки. Відстані до центрів зайнятості Бостона теж не сильно впливає на ціну житла.

Чи буде краще прогнозування ціни, якщо використовувати більш складну модель?

Я на практиці перевірив що так, більш складна модель зробить прогноз більш точним. Якщо порівнювати звичайну регресію і поліноміальну (поліноміальна є більш складною і гнучкою), то остання прогнозує набагато точніше. Але в моїх побудованих моделях є і більш складні моделі, такі як дерева і к найближчих сусідів, вони хоч і прогнозують краще ніж лінійна регресія, але здебільшого програють прогнозам поліноміальної регресії

1. Використані джерела:

<https://www.kaggle.com/vikrishnan/boston-house-prices>

<https://docs.google.com/document/d/1sC5DfBRX8_CHgAwfRo85tDuYJ3BP4veDnXa4U7vQLNw/edit>

<https://docs.google.com/document/d/1icj0t91b4LbBLE0IqqSNTa8lzUjPoEDiONQ801_AY6Q/edit>

<https://www.youtube.com/watch?v=6FouVURil5I&t=2952s&ab_channel=%D0%A2%D0%B0%D1%82%D1%8C%D1%8F%D0%BD%D0%B0%D0%9B%D0%B8%D1%85%D0%BE%D1%83%D0%B7%D0%BE%D0%B2%D0%B0>

<https://scikit-learn.org/stable/modules/cross_validation.html>

<https://scikit-learn.org>

https://slack.com/intl