Софийски университет „Св. Кл. Охридски”

Факултет по математика и информатика

Курсов Проект

Интелигентни системи (Data Mining)

на тема

Обективна оценка на недвижими имоти

Калоян Емилов Николов

Курс: **IV**, Учебна година: **2021/22**

Преподавател: **проф. Иван Койчев**,

Консултант: **ас. Борис Величков**

София, Февруари 2022 г.

Съдържание

1 УВОД 3

1.1 Мотивация 3

1.2 Цел 3

1.3 Ползи 3

2 ПРЕДЛЕГ НА ПОДХОДИТЕ ЗА ОЦЕНКА НА НЕДВИЖИМИ ИМОТИ 4

2.1 Текущо състояние 4

2.2 Избран подход 4

3 НАБОР ОТ ДАННИ 5

4 ИЗПОЛЗВАНИ ТЕХНОЛОГИИ, ПЛАТФОРМИ И БИБЛИОТЕКИ 6

4.1 Технологии 6

4.2 Платформи 6

4.3 Библиотеки 6

5 РЕАЛИЗАЦИЯ 6

6 ТЕСТВАНЕ И ЕКСПЕРИМЕНТИРАНЕ 7

6.1 Агрегираща функция 7

6.2 Брой дървета 8

6.3 Минимален размер на извадката 8

6.4 Максимална дълбочина на дърветата 8

6.5 Минимално понижение в стандартното отклонение 9

6.6 Случаен избор на характеристики 9

6.7 Многократна или еднократна употреба на характеристиките 10

7 ЗАКЛЮЧЕНИЕ 10

8 ИЗПОЛЗВАНА ЛИТЕРАТУРА 11

1 Увод

1.1 Мотивация

Недвижимите имоти – земи и сгради, са важен актив както в живота на отделния човек така и за бизнеса. Определянето на техните цени играе ключова роля във вътрешнофирмените и междуфирмени отношения. Недвижимите имоти са основна част от имуществото на фирмите, хората и държавата. Оценката и преоценката на тяхната стойност носи съществени ползи.

Стойността на недвижимите имоти помага на участниците в стопанския живот да уреждат взаимоотношенията помежду си и да ползват заеми за нуждите на своя бизнес. Хората купуват и продават имоти както с цел за придобиване на доходи, така и за лични нужди. Реалната, достоверна и обективна оценка е от изключително значение за всички тези процеси. Тя защитава и способства интересите както на бизнеса, така и на хората.

1.2 Цел

Настоящият проект има амбицията да създаде инструмент, който не просто да успее да намали субективния човешки фактор, а дори да го изключи напълно при изготвяне на оценка на недвижим имот. Използвайки статистически данни за цените през даден период от време и регион (държава, област, град и т.н.), на базата на множество наблюдения от недвижими имоти, със съответните им характеристики, може да се предвиди стойността на други недвжими имоти от съответния времеви период и географска област. Така лесно и бързо ще се намери коректна и най-важното – обективна оценка на недвижимия имот, която не е обременена от субективното мнение на един или няколко човека.

1.3 Ползи

Подобен инструмент или приложение ще намери употреба в банковия сектор при определяне стойността на недвижим имот, който клиентът желае да използва за обезпечение на кредит. Инструментът ще подпомогне дейността на банковите служители, които без да са специалисти ще получават точна представа за предлаганите обезпечения.

Приложението ще подпомогне и работата на оценителите като премахне нуждата от ръчно събиране на данни за недвижими имоти, сходни с оценявания. Също така може да се използва като коректив на техните оценки, изготвяни чрез прилагане на финансово-счетоводни методи.

Инструментът ще бъде от значителна полза за държавната и данъчна администрация при изготвяне на данъчни ставки за дължимите данъци и при изготвяне на т. нар. данъчна оценка на имотите, която е важен документ при всякакви сделки от страна на бизнеса и гражданите.

Процесът по покупко-продажба на жилище ще стане много по-бърз и прозрачен. Продавачът ще може точно да оцени своя имот и да разбере каква е неговата обективна оценка. Същевременно, купувачите ще могат да проверят дали исканата цена е реалната стойност на съответния имот.

Всички тези причини са достатъчна мотивация за осъществяването на подобен проект. Ако създаденият инструмент е точен и ефективен, той ще намери

приложение в множество области от живота на хората и ще подобри качеството на услугите в редица сектори.

2 Предлег на подходите за оценка на недвижими имоти

2.1 Текущо състояние

Към настоящия момент, оценките на недвижими имоти, съгласно нормативната уредба на България, се извършват от експерти оценители. Изготвяните от тях оценки се използват при кандидатстване за кредит, при покупко-прадажба на имоти и при делба между собственици.

Съгласно „Български Стандарти за Оценяване“ на Камарата на независимите оценители в България[5], основните подходи за оценяване са: приходен подход, разходен подход и сравнителен подход.

При приходния подход, настоящата стойност на имота се сформира от очакваните постъпления от бъдещите парични потоци. Същността на този подход са възможните приходи от обекта като се отчитат бъдещите доходи и спестените разходи. Методи за прилагане на приходния подход са метод на дисконтираните парични потоци и метод на капитализацията.

При разходния подход, стойността на оценявания недвижим имот се изчислява на базата на подробна оценка на необходимите разходи за създаването или придобиване на актива или подобен такъв по предназначение и полезност. Този подход се базира на предположението, че стойността на актива не бива да надвишава разходите по изграждането и придобиването му. Основен метод за прилагането на подхода е методът на амортизираната възстановителна стойност. Той се основава на определяне на всички преки и непреки разходи за изграждане и придобиване на оценявания обект към момента на оценката.

Сравнителният подход дава оценка на имота чрез сравняване със сходни обекти, за които съществуват достатъчни и достоверни знания. Основният метод за прилагането му е метод на пазарните аналози. Оценката се основава на сравнение на количествените и качествените характеристики между оценявания обект и неговите аналози.

2.2 Избран подход

Общият недостатък на всички изброени по-горе подходи за оценка на недвижим имот е, че се базират на финансово-счетоводни изчисления като получените оценки се влияят от субективното мнение на експерта оценител.

Днес, въпреки големите постижения в сферата на информационните технологии и по-специално машинното самообучение, все още когато трябва да се направи оценка на стойност на недвижим имот, то това се извършва от експерт оценител. Ето защо сме си поставили за цел да създадем инструмент, който е способен да прави точна и обективна оценка на недвижими имоти.

Поставената задача изисква да се предвиди цена т.е. някаква числова стойност. Поради тази причина избраният подход включва използването на дървета на регресията. Съгласно източник [3], то дърветата на регресията и като цяло дърветата на решенията са склонни към пренагаждане *(overfitting)* т.е. прекалено добре да научат данните от обучителното множество и съответно да постигат

слаби резултати върху тестовото множество. Ето защо, в рамките на проекта, дървото на регресията се надгражда като се реализира случайна гора *(random forest)*, която стъпва на построяването на множество от дървета на регресията и агрегиране на техните резултати за получаване на крайната оценка.

3 Набор от данни

Наборът от данни, който се използва в рамките на проекта е взет от Kaggle. Той е създаден и предоставен от Тони Пино. Съдържа данни за цените и характеристиките на недвижими имоти в град Мелбърн, Австралия към септември 2017 г.

За целите на проекта е извършена предварителна обработка на данните, при която някои от характеристиките са премахнати. Адрес, Географска ширина, Географска дължина и Дата на осъществяване на продажбата са премахнати, тъй като притежават уникални стойности за всеки недвижим имот. Пощенски код също е премахнат, тъй като се оказва, че стойностите на тази характеристика са свързани с тези на „Квартал”.

Наборът от данни, който е използван при получаването на описаните в Секция 6. резултати, включва следните характеристики на недвижимите имоти:

- Предградие;
- Брой стаи;
- Тип на недвижимия имот:
    - h – house – къща, която е отделена от други недвижими имоти;
    - t – townhouse – къща, чиито стени се допират до други къщи, често със сходна архитектура;
    - u – unit – апартамент в къща (duplex), която има 2 отделни входа за 2 отделни апартамента.
- Брокер на недвижимия имот;
- Брой спални;
- Брой бани;
- Брой паркоместа;
- Площ в кв. м.
- Година на построяване;
- Квартал;
- Район;
- Цена.

Melb-data.csv съдържа 6160 наблюдения и тук с цел онагледяване са представени първите 5 от тях:

Предградие	Стаи	Тип	Брокер	Спални	Бани	Парко-места	Площ	Год.	Квар-тал	Район	Цена
Abbotsford	2	h	Biggin	2	1	0	156	1900	Yarra	Northern Metropolitan	1035000
Abbotsford	3	h	Biggin	3	2	0	134	1900	Yarra	Northern Metropolitan	1465000
Abbotsford	4	h	Nelson	3	1	2	120	2014	Yarra	Northern Metropolitan	1600000
Abbotsford	3	h	Nelson	4	2	0	245	1910	Yarra	Northern Metropolitan	1876000
Abbotsford	2	h	Nelson	2	1	2	256	1890	Yarra	Northern Metropolitan	1636000

Таблица 1. Съдържание на Melb-data.csv

4 Използвани технологии, платформи и библиотеки

4.1 Технологии

Програмният език, който е използван е C++.

C++ се използва за реализиране на приложения, при които трябва да се използва максимално ефективно наличния хардуер. Приложения, разработени на C++ постигат висока производителност. Езикът включва необходимите инструменти за възползване от ползите както на хоризонтално, така и на вертикално мащабиране.

Същевременно, C++ има и някои недостатъци, които винаги затрудняват и забавят процеса на разработка – нужда от използване на указатели, динамично заделяне на памет и други[4].

4.2 Платформи

C++ е език, който изисква компилация. Това означава, че е необходим C++ компилатор, който да генерира съответния изпълним файл, в зависимост от използваната платформа. При условие, че е наличен подходящ компилатор, приложението може да се изпълни на всяка платформа.

4.3 Библиотеки

За реализацията на проекта **не** са използвани специализирани библиотеки. Дървото на регресията, както и случайната гора са имплементирани от нулата като част от процеса по разработка на проекта.

5 Реализация

Сорс кода на проекта е разделен по следния начин:

Header Files:
- Bootstrap.h
- Node.h
- RegressionTree.h
- RandomForest.h
- Tests.h

Source Files:
- Bootstrap.cpp
- Node.cpp
- RegressionTree.cpp
- RandomForest.cpp
- Tests.cpp
- Main.cpp

Bootstrap.h и Boopstrap.cpp декларират и дефинират логиката за прочитане на набора от данни, неговата предварителна обработка и различни начини за формиране на извадки от набора от данни.

Node.h и Node.cpp отговарят за това как се представят върховете в паметта, както и помощни функции за инициализиране, достъпване и промяна на член данните на тези обекти.

RegressionTree.h и RegressionTree.cpp декларират и дефинират цялостната логика по построяване на дърво на регресията, както и проверка на неговата точност.

RandomForest.h и RandomForest.cpp съдържат логиката за построяването на случайна гора от дървета на регресията (обекти от тип RegressionTree), както и тестване на нейните резултати.

Tests.h и Tests.cpp съдържат тестове за проверка точността на случайната гора при различни стойности на използваните параметри.

6 Тестване и експериментиране

6.1 Агрегираща функция

В рамките на проекта са сравнени 4 различни агрегиращи функции.

1. Средна стойност – крайната оценка на случайната гора е средно аритметично от оценките на всички дървета на регресията.
2. Медиана – крайната оценка на случайната гора е медианата от оценките на всички дървета на регресията.
3. Функция, при която се отегляват линейно оценките на различните дървета на регресията в зависимост от тяхната точност. По-конкретно – ако например има само 2 дървета и едното (tree1) е с 50% по-висока точност от другото (tree2), то крайната оценка ще се получава по формулата:

   *predictedPrice = 0.6 * tree1_predictedPrice + 0.4 * tree2_predictedPrice*
4. Функция, при която се отегляват оценките на различните дървета на регресията в зависимост от тяхната точност, повдигната на степен X. В представените по-долу резултати X = 4. В този случай, ако имаме 2 дървета и едното (tree1) е с 50% по-висока точност от другото (tree2), то крайната оценка ще се получава по формулата:

   *predictedPrice = 0.835 * tree1_predictedPrice + 0.165 * tree2_predictedPrice*

При Агрегиращи функции 3 и 4, обучителното множество се използва за генериране на извадки, върху които се обучават дърветата на регресията. Валидиращо множество се използва за намиране на точността на всяко от дърветата и съответното му тегло за крайната оценка. Тестовото множество се използва за проверка точността на случайната гора.

Тъй като се определя числова стойност, то за оценка на точността на алгоритъма, не можем да използваме колко пъти предвидената стойност съвпада с реалната. Ето защо в Таблица 2. са използвани критерии като корен квадратен от средната квадратична грешка *(Root Mean Square Error)*[2] и процента наблюдения от тестовото множество, оценени коректно с грешка не повече от 10%, 25%, 50% и 75%.

Агрегираща функция	Средна Грешка*	-10 < x < 10	-25 < x < 25	-50 < x < 50	-75 < x < 75
1	478842	21.5	50.2	75.3	86.8
2	454037	23.7	53.9	80.0	89.2
3	468099	22.9	51.5	76.7	87.5
4	439815	24.5	54.7	78.9	88.9

* Корен квадратен от средната квадратична грешка

Таблица 2. Резултати в проценти при различни агрегиращи функции

Както виждаме от таблица 2, Агрегираща функция 4 се представя най-добре и затова в следващите експерименти се използват постигнатите резултати от нея.

6.2 Брой дървета

В Таблица 3. са показани постигнатите резултати при различен брой дървета:

Брой дървета	Средна Грешка*	-10 < x < 10	-25 < x < 25	-50 < x < 50	-75 < x < 75
2	620666	17.9	44.7	71.7	83.3
4	549538	22.8	53.4	78.5	88.8
6	576462	21.7	52.5	77.1	87.5
8	516042	23.7	52.3	77.9	89.1
10	543906	23.0	52.8	78.6	86.3
20	512193	24.2	53.7	79.5	89.4
30	509008	25.1	53.8	78.7	89.4
50	498705	25.2	55.1	81.3	90.0
75	521706	23.7	53.6	77.9	88.8
100	514934	23.7	54.0	78.2	88.4
200	538013	22.4	53.2	78.5	88.3

* Корен квадратен от средната квадратична грешка

Таблица 3. Резултати в проценти при различен брой дървета

От представените резултати се забелязва, че точността на случайната гора е приблизително сходна при 8 или повече дървета на регресията. Ако се изследват внимателно резултатите, се забелязва, че най-добра точност е постигната при 50 дървета на регресията.

6.3 Минимален размер на извадката

В Таблица 4. са представени резултатите при различен размер на минималната извадка за строене на дървото на регресията. Когато броя наблюдения за даден връх е под минималния, то върхът не се разделя, а се определя за листо[1]. За получаването на представените резултати, всяко дърво се обучава и съответно построява на база на 2000 случайно избрани примера от обучителното множество.

Мин. размер на извадката	Средна Грешка*	-10 < x < 10	-25 < x < 25	-50 < x < 50	-75 < x < 75
1	512306	22.6	51.5	75.6	86.0
5	491136	23.1	52.8	80.0	89.1
10	484672	25.0	53.2	79.1	89.9
50	462851	26.8	55.5	80.1	90.3
75	481211	25.8	55.2	81.0	89.6
100	490539	24.4	53.8	80.6	90.0
150	498627	23.8	52.4	80.1	89.6
200	512016	23.6	51.6	79.7	89.4
300	515383	22.9	52.9	78.9	88.7
500	512989	23.1	52.4	79.0	88.3

* Корен квадратен от средната квадратична грешка

Таблица 4. Резултати в проценти при различен минимален размер на извадката.

Съгласно представените резултати, случайната гора постига най-голяма точност и съответно минимална грешка при размер на минималната извадка при строене на дърветата от около 50 наблюдения.

6.4 Максимална дълбочина на дърветата

В Таблица 5. са показани резултатите при различна максимална дълбочина на изградените дървета на регресията. При дефинирането на дълбочината се приема, че коренът е на ниво (дълбочина) 1 и поради тази причина, за да се изгради дърво, е необходимо максималната дълбочина да бъде поне 2.

Максимална дълбочина	Средна Грешка*	-10 < x < 10	-25 < x < 25	-50 < x < 50	-75 < x < 75
2	539734	21.7	51.4	72.7	84.2
3	516794	22.9	52.3	76.8	88.2
5	483180	24.6	53.8	79.3	90.0
10	485965	24.5	53.6	79.5	90.0
15	488506	24.3	53.3	78.9	89.0
20	492654	25.3	53.9	79.6	89.2
25	501654	24.2	52.7	76.2	88.7
30	512811	23.1	52.1	74.6	87.0

* Корен квадратен от средната квадратична грешка

Таблица 5. Резултати в проценти при различна максимална дълбочина на дърветата.

От представените резултати се вижда, че случайната гора постига най-добра точност при максимална дълбочина на дърветата на регресията 20. Въпреки това, трябва да се отбележи, че на практика разликите в интервала между максимална дълбочина 5 и 20 са пренебрежимо малки.

6.5 Минимално понижение в стандартното отклонение

В Таблица 6. са представени резултати при различни стойности на минимално намаление на стандартното отклонение. В процеса на изграждане на всяко дърво на регресията, когато трябва да се избере атрибут за съответния връх, се избира този атрубит (със съответната точка на делене), който ще донесе най-голямо понижение на стандартното отклонение. Ако това понижение е по-малко от минималното, което се изисква, то съответният връх се определя за листо.

Минимално понижение**	Средна Грешка*	-10 < x < 10	-25 < x < 25	-50 < x < 50	-75 < x < 75
1	489617	24.2	53.4	76.8	88.5
5	476259	26.0	56.8	79.4	90.3
10	470168	26.2	57.1	80.1	90.1
20	470168	26.2	57.1	80.1	90.1
50	461053	27.7	57.2	80.1	90.0
100	483659	27.4	57.5	79.5	89.9
200	491792	27.3	56.8	80.1	89.9
500	493662	26.0	57.0	79.7	89.5
1000	510659	25.8	56.7	78.9	89.4
2000	515812	25.6	56.4	78.6	89.2

* Корен квадратен от средната квадратична грешка

** Минимално понижение на стандартното отклонение

Таблица 6. Резултати в проценти при различни стойности на минималното понижение на стандартното отклонение.

Съгласно представените резултати, най-добра точност се постига при минимално понижение на стандартното отклонение 50 и поради тази причина в следващите изследвания ще бъде използвана тази стойност.

6.6 Случаен избор на характеристики

Съгласно източник [3], при изграждането на дърветата на регресията трябва при избор на атрибут или характеристика за всеки от върховете, да се избира измежду само няколко от наличните характеристики на наблюденията. Тези няколко характеристики трябва да се избират на случаен принцип и да са различни за всеки връх. В тази секция се изследва разликата в представянето на алгоритъма при 2 различни начина за определяне между кои характеристики да се избира за всеки връх:

1. Характеристики се избират на случаен принцип за всяко дърво. За всеки връх в рамките на едно и също дърво се избира от едно и също подмножество от наличните характеристики.
2. Характеристики се избират на случаен принцип за всеки връх. За всеки връх от всяко дърво се избира на случаен принцип ново подмножество от наличните характеристики. Разбира се, макар и малко вероятно, отново е възможно 2 върха да избират измежду едно и също подмножество от характеристики.

В Таблица 7. са представени наблюдаваните резултати:

Подход	Средна Грешка*	-10 < x < 10	-25 < x < 25	-50 < x < 50	-75 < x < 75
За всяко дърво	537208	23.4	49.8	76.7	87.6
За всеки връх	461053	27.7	57.2	80.1	90.0

* Корен квадратен от средната квадратична грешка

Таблица 7. Резултати в проценти при различни подходи за избор на характеристики.

Както се вижда от Таблица 7., значително по-добра точност се постига, ако се използва втория подход – да се избира подмножество от характеристики за всеки връх за всяко дърво, а не само за всяко дърво.

6.7 Многократна или еднократна употреба на характеристиките

В тази секция се сравняват точността на алгоритъма в зависимист от това дали е възможно да се използват характеристиките повече от веднъж. Например, ако характеристиката „Брой стаи“ може да се използва само веднъж, но данните ще се разделят напрамер в засимист от това дали стаите са повече от 3 или не. Ако е възможно да се използва тази характеристика многократно, то е възможно да се образуват интервали, спрямо които да се разделят наблюденията. Например брой стаи $Y \in [1,3]$, $Y \in [4,5]$ и $Y \in [6, \infty]$.

В Таблица 8. са представени наблюдаваните резултати:

Употреба	Средна Грешка*	-10 < x < 10	-25 < x < 25	-50 < x < 50	-75 < x < 75
Еднократна	522524	24.4	50.3	78.9	89.3
Многократна	461053	27.7	57.2	80.1	90.0

* Корен квадратен от средната квадратична грешка

Таблица 8. Резултати в проценти при еднократна и многократна употреба на характеристиките.

От получените резултати, следва, че случайната гора постига по-висока точност, ако е разрешена многократната употреба на характиристики.

7 Заключение

В рамките на проекта не просто се реализира случайна гора от дървета на регресията с цел обективна оценка на недвижими имоти, но и се изследваха различни нейни параметри с цел подобряване на наблюдаваните резултати. Сред параметрите, чиито стойности бяха изследвани са: брой дървета на регресията, които се обучават; минимален размер на извадката при изграждане на дърветата; максимална дълбочина

на дърветата; минимално понижение в стандартното отклонение, което ни носи избрана характеристика; начин за подбор на характеристики; дали те да се използват еднократно или многократно, както и са сравнени резултатите при четири различни агрегиращи функции.

По-нататъшното развие на проекта включва тестване на представянето на инструмента и върху други набори от данни – съдържащи наблюдения на недвижими имоти в Европа и дори България. Също така, може да се разшири насочеността на приложението и към други типове недвижими имоти, тъй като към момента в използвания набор от данни се използват само няколко типа.

8 Използвана литература

[1] Jason Brownlee, Classification And Regression Trees for Machine Learning, published: April 8, 2016, last updated: August 15, 2020, https://machinelearningmastery.com/classification-and-regression-trees-for-machine-learning/

[2] Ashwin Prasad, Regression Trees | Decision Tree for Regression | Machine Learning, published: Aug 8, 2021, https://medium.com/analytics-vidhya/regression-trees-decision-tree-for-regression-machine-learning-e4d7525d8047

[3] Michelle Jane Tat, Seeing the random forest from the decision trees: An explanation of Random Forest, published: April 16, 2017, https://towardsdatascience.com/seeing-the-random-forest-from-the-decision-trees-an-intuitive-explanation-of-random-forest-beaa2d6a0d80

[4] Advantages and Disadvantages of C++, https://techvidvan.com/tutorials/cpp-pros-and-cons/

[5] Камара на независимите оценители в България, Български стандарти за оценяване, 2018, https://private.ciab-bg.com/uploads/common/hkxfb5j0pm19g2ya.pdf

[6] Данните са взети от https://www.kaggle.com/code/dansbecker/random-forests/data?select=melb_data.csv