ЦЕНЕ КУЋА: НАПРЕДНЕ ТЕХНИКЕ РЕГРЕСИЈЕ

Овим документом дат је предлог пројекта из предмета Системи за истраживање и анализу пројекта. Дата је дефиниција проблема и објашњен је допринос и значај његовог решавања. Наведени су, и укратко описани, радови који се баве решавањем сродних проблема. Описани су скуп података, методологија решавања проблема, као и метод евалуације. Наведен је софтвер који ће бити коришћен.

ДЕФИНИЦИЈА ПРОБЛЕМА

Предикција цена кућа на основу њихових аспеката изражених путем 79 одабраних варијабли. Идеја пројекта је олакшавање процене вредности дате куће на основу одређеног броја различитих критеријума.

МОТИВАЦИЈА

У питању је такмичарски задатак који подстиче учеснике да одреде како неке карактеристике кућа, које можда на први поглед и немају директне везе са процењеном ценом куће, утичу на њену вредност.

Заинтересованих страна за решавање овог проблема је много, а разлог је чисто економске природе. Поред учесника на тржишту некретнина, заинтересовани су и држава, ради боље наплате пореза, осигуравајућа друштва, ради реалније процене вредности некретнине, и други.

РЕЛЕВАНТНА ЛИТЕРАТУРА

[1] Limsombunchao, V. (2004). House price prediction: Hedonic price model vs. artificial neural network

Тема рада: Емпиријско поређење успешности хедоничког регресионог модела у одосу на вештачке неуронске мреже приликом предикције цене кућа..

Методологије: За поређење су кориштени WLS (Weighted Least Squares) регресиони модел и вештачка неуронска мрежа (feed-forward/back-propagation). Предикција се врши у зависности од појединачних карактеристика кућа

Подаци: Прикупљени су подаци о 200 кућа Кристчрч (Christchurch) области у Новом Зеланду. Подаци обухватају карактеристике као што су величина и старост куће, број спаваћих соба и купатила, географска локација, итд.

Евалуација решења: Извршена је насумична подела података – 80% података је кориштено за обуку, док је осталих 20% кориштено за тестирање. Потом су израчунате R^2 и RMSE вредности, касније употребљене за поређење.

Закључак: Вештачка неуронска мрежа се показала као прецизнија метода, за дати скуп података.

Коментар: Мана је релативно мали скуп тест података.

[2] Feng, X., & Humphreys, B. (2016). Assessing the Economic Impact of Sports Facilities on Residential Property Values A Spatial Hedonic Approach. Journal of Sports Economics, 1527002515622318.

Тема рада: Доказати да близина кућа спортским објектима позитивно утиче на њихову цену.

Методологије: Коришћене су метода најмањих квадрата, метод процене највеће вероватноће (Maximum Likelihood Estimation), и двофазна метода најмањег квадрата (која узима у обзир просторну зависност - *spatial*).

Подаци: Прикупљени су подаци о 9504 кућа, продатих 2000-те године, смештених у Колумбусу, Охајо. Подаци обухватају карактеристике самих кућа, као што су број спаваћих соба, постојање подрума и гараже, али такође обухватају и карактеристике насеља као што је средњи доходак, структура становништва и квалитет школа. Затим, укључене су и удаљености насеља од већих транспортних рута, предузетничке четврти и од два спортска објекта.

Евалуација решења: Израчунате су R^2 вредности сваког од модела. Метода најмањег квадрата се показала као најнеуспешнија, док се најуспешнијом показала двофазна метода најмањег квадрата (која узима у обзир просторну зависност - spatial).

Закључак: Доказана је корелација између удаљености куће од спортског центра и цене куће – што је објекат ближе спортском центру, то му је већа цена (пад цене од 0.14% приликом повећања раздаљине за 1%).

[3] Bin, O., Kruse, J. B., Landry, C. E.. (2008). Flood Hazards, Insurance Rates, and Amenities: Evidence From the Coastal Housing Market. The Journal of Risk and Insurance

Тема рада: Утицај ризика од поплава на вредности приобалних некретнина.

Методологије: Коришћена је линеарна регресија.

Подаци: Прикупљени су подаци о продаји некретнина у Картарет каунтију (Cartaret County) између 2000. и 2004. године. Картарет каунти се налази на обали Атлантског океана у Северној Каролини. У највишој тачки се налази на 15 метара надморске висине и склон је поплавама, тако да је био погодан за истраживање. Због инфлације, све цене су нормализоване на јачину америчког долара 2004. године. Осим цена, подаци обухватају карактеристике некретнина, као што су старост, величина објекта, величина плаца, број купатила, удаљеност од обале. По подложности поплавама, некретнине су подељене у три категорије: оне које нису подложне поплавама, оне које имају 0.2% шансе годишње да буду поплављене, и оне које имају 1% годишњи ризик од поплаве.

Евалуација решења: Направљена су четири модела, од којих прва два нису правила дистинкцију између две категорије подложне поплавама, док друга два јесу. Први и трећи модел нису узимали у обзир удаљеност од обале, за разлику од другог и четвртог. Рачунате су χ^2 вредности. Четврти модел је узет као најрепрезентативнији.

Закључак: Студија указује да је некретнина која се налази у поплавном подручју значајно јефтинија од некретнине сличних карактеристика изван њега. Поплавно подручје умањује процењену вредност некретнине просечно за $11,598, што представља снижење од просечно 7,3%. Разлика у цени између подручја са већим и мањим ризиком од поплава је, такође, статистиички значајна, али се може приметити да се за пет пута већи ризик од поплава бележи пад цене од само 30%.

ОСТАЛЕ РЕФЕРЕНЦЕ

[4] Рад у ком је конструисан скуп података - https://ww2.amstat.org/publications/jse/v19n3/decock.pdf (додато 30.11.2016. године).

[5] Адреса на којој се могу наћи и тренинг и тестни скуп података - https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data (додато 30.11.2016. године).

[6] Детаљан опис скупа података - https://kaggle2.blob.core.windows.net/competitions-data/kaggle/5407/data_description.txt?sv=2015-12-11&sr=b&sig=TU45lmGDOQfanmHMEdY1qHsUiCIRbOIT5L19fV7fE1U%3D&se=2016-12-02T16%3A10%3A03Z&sp=r (додато 30.11.2016. године).

[7] Метод евалуације, прописан од стране организатора такмичења - https://www.kaggle.com/c/house-prices-advanced-regression-techniques/details/evaluation (додато 30.11.2016. године).

СКУП ПОДАТАКА

Скуп података креиран је као проширена и модернизована алтернатива познатом *Boston Housing* скупу података. Рад у ком је креиран скуп података доступан је у [4].

Скуп се састоји од 2919 кућа са њима припадајућим обележјима, при чему је од стране организатора такмичења он подељен на 2 дисјунктна подскупа, у односу 50% на према 50%, на тренинг и тестни подскуп података (оба се могу наћи на [5]). Свака кућа је описана са 79 обележја, попут свеукупног стања куће, облика плаца, врсте крова итд. Продајна цена куће, која је заправо зависна променљива која се одређује на основу осталих обележја, дата је само у тренинг скупу података. Детаљан опис већине обележја могуће је пронаћи у [6].

МЕТОДОЛОГИЈА

За решавање проблема биће искоришћене следеће методе:

- линеарна, *Ridge*, *Lasso* и *Elastic Nets* регресија;
- вештачке неуронске мреже;
- случајне шуме (*Random Forests).*

СОФТВЕР

Пројекат ће бити реализован коришћењем *R* програмског језика, у оквиру *Visual Studio 2015* развојног окружења. За вршење експлоративне анализе биће коришћен софтверски алат *RapidMiner Studio 7*.

МЕТОД ЕВАЛУАЦИЈЕ

С обзиром на то да се ради о такмичарском задатку, метод евалуације је унапред одређен (видети [7]). Резултати се евалуирају коришћењем *Root-Mean-Squared-Error* методе између логаритма предвиђене вредности и логаритма стварне вредности. Логаритмовањем ових вредности постиже се то да предикције за скупе куће и предикције за јефтине куће једнако утичу на укупни резултат. Евалуација се врши над резултатима добијеним применом модела над тестним скупом података, издвојеним од стране организатора такмичења.

ПЛАН

Реализација пројекта би требало да обухвати следеће прекретнице (*milestones*):

- експлоративна анализа, кластер анализа (*Model Based Clustering*), визуализација модела и визуализација података (*PCA* или *MDS*),
- претпроцесирање података (нормализација, надомешћивање непознатих вредности, додавање додатних променљивих за категоричне вредности и слично),
- креирање модела на основу одабраних методологија,
- евалуација модела и прилагођавање параметара и
- анализа добијених резултата.

Comment [WU1]: http://www.stat.washington.edu/research/reports/2012/tr597.pdf

ТИМ

Тим чине следећи чланови: Ненад Тодоровић (E2 33/2012), Никола Тодоровић (E2 31/2016) и Давид Вулетић (E2 79/2016).