ივანე ჯავახიშვილის სახელობის თბილისის სახელმწიფო უნივერსიტეტი
ზუსტ და საბუნებისმეტყველო მეცნიერებათა ფაკულტეტი

ჯგუფური პროექტი

უძრავი ქონების ფასების პროგნოზირება მანქანური სწავლების საშუალებით

საბაკალავრო პროგრამა - კომპიუტერული მეცნიერება

პროექტის შემსრულებლები:
დავით კობალაძე
გიორგი ბერიძე

ხელმძღვანელი: ასისტენტ-პროფესორი, ლექტორი,
არჩუაძე მაია

თბილისი
2019

ანოტაცია

მოცემული პროექტის მიზანია გავამარტივოთ ადამიანების ცხოვრება და მივცეთ საშუალება ამ კონკრეტულ მაგალითზე, სახლის ყიდვის ან გაყიდვის საშუალება მარტივად და სარფიანად. ჩვენი პროექტის მიზანია კლიენტმა მარტივად შეძლოს საკუთარი სახლის (გაყიდვის შემთხვევაში) ფასის დაანგარიშება და მას არ დაჭირდეს ამისთვის ზედმეტი ხარჯის გაღება და ვინმეს დაქირავება რომ მას დაეხმაროს სახლის გაყიდვაში (შეიძლება ამ დროს მოტყუებულიც დარჩეს). ჩვენი პროგრამა მარტივი გამოსაყენებელია და არ მოითხოვ დიდი დროს და ენერეგიას, უბრალოთ მარტო საჭირო იქნება სახლოს მონაცემების განთავსება ჩვენს პროგრამაში და სურ რამდენიმე წამში თქვენ შეძლებთ გაიგოთ თქვენი სახლის ფასი, რაც დაგეხმარებად არ მოტყუვდეთ და მარტივად გაყიდოთ სახლი. ასევე დაგეხმარებათ სახლის შეძენი დროს, მარტივად დაადგენ ფასს, რომელი სახლის შეძენასაც გადაწყვეტთ და არ გექნებათ ეჭვი მეტი ხომ არ გიწევთ გადახდა.

მონაცემების მოგროვება აქტიური კვლევის საკითხია მანქანური სწავლების საზოგადოებებში. არსებობს ორი უმთავრესი მიზეზი რის გამოც მონაცემების მოგროვება გახდა კრიტიკული საკითხი. პირველი ისაა რომ მანქანური სწავლის მეთოდების გამოყენება დაიწყეს უფრო ფართოდ, ამ მეთოდის გამოყენება დაიწყო ისეთ სფეროებში სადაც არაა საკმარისი რაოდენობის დახარისხებული მონაცემები. მეორე მიზეზი ისაა რომ ტრადიციული მანქანური სწავლებისგან განსხვავებით როდესაც თვისებების აგება (Feature Engineering) ხდებოდა მექანიკურად, ღრმა სწავლების მეთოდებს თავად შეუძლიათ თვისებების დასწავლა მაგრამ საამისოდ ჭირდებათ ბევრად მეტი დახარისხებული ინფორმაცია. მანქანური სწავლების ალგორითმისთვის მონაცემების მოგროვება მოიცავს უშუალოდ მონაცემების მოძიებას, ამ მონაცემების დახარისხებას და არსებული მონაცემების გაუმჯობესებას. ჩვენ შევეცადეთ გვეჩვენებინა ვებ-საიტებიდან მონაცემების მოგროვების ავტომატური მეთოდების ეფექტურობა უძრავი ქონების ვებ-საიტების მაგალითზე. ჩვენ შევძელით 200 000 ჩანაწერის მოგროვება 3 ვებ-საიტიდან რაც საკმარისი აღმოჩნდა მანქანური სწავლების მეთოდებით პროგნოზირებისას საკმაოდ მაღალი სიზუსტის მისაღწევად.

Predicting Real Estate Prices Using Machine Learning

Abstract

We created system for automatizing data collection. We've shown efficiency of this method in conjunction with Machine Learning methods on predicting real estate prices. To make this system accurate we needed a lot of data to train model on it. We collected housing prices from three most popular Georgian real estate web-sites: ss.ge, myhome.ge, place.ge.

შინაარსი

შესავალი

ელხა ვისაუბრებ თუ საიდან წამოვიდა ეს იდეა, ჩვენ დაინტერესებული ვიყავით მანქანური სწავლებით, ამ სფეროში გვქონდა თეორიული ცოდნა, შესავამისად გვინდოდა ამის პრაქტიკაში გამოყენება, დავიწყეთ ფიქრი თუ რა შეგვეძლო გაგვეკეთებინა, რითაც ადამიანის ცხოვრებას გავამარტივებდი. როგორც მოგეხსენებათ მანქანურ სწავლებას ჭირდება დიდი მონაცემები, მონაცემების სიმცირის ან საერთოდ არ ქონის გამო ბევრ იდეაზე ვთქვით უარი, შემდეგ გადავწყვიტეთ ჯერ გვენახა ისეთი რამ რომლის გასაწვთნელად საკმარის მონაცემების შეგროვებას შევძლებდით, ჩვენი ძიების შემდეგ მივედი იქამდე რომ მონაცემის შეგროვება შეგვეძლო ონლაინ გაყიდვების საიტებიდან, მაგრამ რაში გამოგვადგებოდა ეს? შემდეგ გავიფიქრეთ თუ რატო არ შეიძლება იყოს ისეთე ტექნოლოგია რომელიც შეძლებს ნივთის ფასის პროგნოზირებას, რის შესახებაც მას ექნება მონაცემები. გავაკეთეთ გამოკითხვა სხვადასხვა ნივთების ფასის ცოდნის შესახებ ხალხში, გამოკითხვის შედეგად მოსახლეობის 30% თუ ერკვევა სახლის ღირებულებაში, ამ მდგომარეობის გამო დავიწყეთ ფიქრის სახლის ფასის პროგნოზირებაზე.

მოდი ეხლა მოკლედ ვისაუბრებ მანქანური სწავლების შესახებ, რატომ გახდა მანქანური სწავლება საჭირო პროგრამირებაში? რამდენად გავცელებულია დღეს მანქანური სწავლების გამოყენებადობა? აქვს თუ არა მანქანურ სწავლებას მომავალი?

შევეცდები მე რამდენიმე სიტყვით ამ კითხვებზე გავცე პასუხი და ავხსნა მანქანური სწავლების საჭიროება დღევანდელ ცხოვრებაში. როდესაც უკვე რთულია ან შეუძლებელია ამოცანის ამოხსნა,პროგრამისტისთვის რთულია და თითქმის შეუძლებელი რომ განიხილოს ყველა ნიუნანსი. მანქანური სწავლებით ეს პრობლემა მარტივად მოგვარებადია, სწავლების ალგორიტმებით, მონაცემების საკმარისი რაოდენობით და მანქანური სწავლების მეთოდით ჩვენ არ გვჭირდება ყველა ნიუანსის განხილვა.

მანქანური სწავლები გამოყენება დღესდღეობით წინ არის წამოწეული და ფარდოდ გამოყენებადია, მას იყენებენ ისეთი პოპილარული კომპანიებიც როგორიცაა facebook , google ,Microsoft და სხვა ბევრი ცნობილი კომპანია.

ჩემი აზრით და ცოდნით მანქანური სწავლება დღითიდღე უფრო და უფრო მნიშვნელოვანი ხდება ჩვენს ცხოვრებაში, თითქმის ყველა მსხვილი კომპანია იყენებს მანქანუს სწავლებას.

ჩვენ ვცხოვრობთ პერიოდში სადაც მანქანურ სწავლებას ახდენს ღრმა გავლენას ბევრ სფეროზე მისი გამოყენების ვრცელი არელით როგორიცაა ტექსტის ანალიზი, ხმის და სურათების ამოცნობა და მრავალი სხვა. გასაოცარი მაგალითია ის რომ ღრმა სწავლების მეთოდები იგივე სიზუსტით ცნობენ დიამეტთან დაკავშირებულ თვალის პრობლემებს სურათებიდან რითაც ოფთალმოლოგები. ამ წარმატებაში უდიდესი წვლილი მიუძღვის გაუმჯობესებულ გამოთვლით ინფრასტრუქტურას და დიდი ოდენობით დახარისხებულ საწვრთნელ მონაცემებს.

მანქანურ სწავლებაში არსებული გამოწვევებიდან მონაცემების მოგროვება და დახარისხება უფრო და უფრო კრიტიკული ხდება. ცნობილია რომ მანქანური სწავლების მოდელის შექმნის ყოველ იტერაციაში ყველა დიდ დროს მონაცემების მოგროვება, გაწმენდა და მისი გააზრება იკავებს. ზემოთ ჩამოთვლილი ყოველი ნაბიჯი დროის წამღებია მაგრამ მონაცემების მოგროვება გახდა ყველაზე კრიტიკული გამოწვევა ქვემოთ მოყვანილი მიზეზების გამო.

პირველი არის ის რომ მანქანური სწავლების ალგორითმების გამოყენება დაიწყეს ისეთ სფეროებში სადაც ან საერთოდ არ არის დახარისხებული მონაცემები ან ამ მონაცემების რაოდენობა ძალზედ მწირია. ტრადიციულ გამოყენებებში, როგორიცაა სურათებში ობიექტების ამოცნობა და ტექსტის თარგმნა, დიდი ოდენობითაა საწვრთნელი მონაცემები რასაც ვერ ვიტყვით სხვა სფეროებზე. მაგალითად დიდი მოთხოვნაა ქარხნებში ხარისხის კონტროლზე მანქანური სწავლების მეთოდების გამოყენებით მაგრამ არ არსებობს თითქმის არანაერი საწვრთნელი მონაცემები. ყოველ ახალ დეტალზე და ნაწილზე ხელით ხდება საჭირო

მონაცემების მოგროვება.

მეორე მიზეზი არის ღრმა სწავლების პოპულარიზაცია რამაც გამოიწვია დიდი ოდენობით მონაცემების არსებობის საჭიროება. ეს განაპირობა იმან რომ ღრმა სწავლების მეთოდები მონაცემებიდან თვისებებს თავად სწავლობენ ტრადიციული მეთოდებისგან განსხვავებით რომელშიც ალგორითმებისთვის თვისებებს მექანიკურად აგებდნენ. ეს თვისებების შესწავლა კი იწვევს იმას რომ კარგი სიზუსტის მისაღწევად ღრმა სწავლების მეთოდებს ბევრად მეტი მონაცემი სჭირდებათ ვიდრე ტრადიციული მანქანური სწავლების მეთოდებს.

ამის შედეგად გაჩნდა დიდი მოთხოვნა მონაცემების მოგროვების მასშტაბურ და ზუსტ მეთოდებზე. მანქანური სწავლების ალგორითმებისთვის მონაცემების მოგროვებისათვის არსებობს ძირითადად სამი მეთოდი: მონაცემების შეგროვება ნულიდან, არსებული მონაცემების გადაკეთება და სინთეტიკური მონაცემების გენერირება. ჩვენი ამოცანის გასაჭრელად ყველაზე მართებული მონაცემების შეგროვება იყო რადგან ქართული ბაზრისთვის არ არსებობს უძრავის ქონების ფასების შესახებ მონაცემები იმ სახით რაც საჭიროა მანქანური სწავლების ალგორითმისათვის.

ჩვენ შევძელით შეგვექმნა პროგრამა რომელიც ავტომატურად აგროვებს მონაცემებს 3 უძრავის ქონების გაყიდვის საიტიდან რომლის მეშვეობითაც რამდენიმე დღის განმავლობაში მოვაგროვეთ 200 000 ჩანაწერი.

არსებული მდგომარეობა

პროგნოზირება მრავალი წელია აქტიური კვლევის საკითხია. პროგნოზირება მრავალ სფეროს მოიცავს და მრავალი გამოყენება აქვს. პროგნოზირებისთვის იყენებენ მრავალ მეთოდს დაწყებული სტატისტიკური მოდელირებით და დამთავრებული მანქანური სწავლების მეთოდებით.

ჩვენ გადავწყვიტეთ გვეჩვენებინა მანქანური სწავლების მეთოდების ეფექტურება უძრავი ქონების ფასების პროგნოზირების მაგალითზე. გადავწყვიტეთ ეს გვეჩვენებინა ქართული უძრავი ქონების ბაზრის მაგალითზე რადგან ამის მაგალითი ქართული ბაზრისთვის არ არსებობს.

ამჟამად ფასების დასადგენად და პროგნოზირებისთვის გამოიყენება ადამიანის ექსპერტიზა რაც ხშირ შემთხვევაში ძვირი და არაზუსტია რადგან ადამიანი ყველაფერს ცივი გონებით ვერ განსჯის.

პროექტის ძირითადი ამოცანის დასმა

ჩვენი უპირველესი ამოცანა იყო მონაცემების მოგროვება რადგან ქართულ ბაზარზე უძრავი ქონების ფასების მონაცემთა ბაზა არ არსებობს.

მოცემული ამოცანის აზრი მდგომარეობს იმაში რომ შეგვექმნა მეთოდი რომლის მეშვებიოთაც შედარებით მცირე დროში მარტივად მოვაგროვებდით დიდი ოდენობის საწვრთნელ მონაცემებს. მიღებული პროგრამული უზრუნველყოფა უნდა იყოს მარტივად სკალირებადი, უნდა შეიძლებოდეს მონაცემების ახალი წყაროების დამატება მინიმალური ცვლილებით კოდში.

შემდეგი მიზანი იყო მანქანური სწავლების მეთოდების მეშვეობით ფასების პროგნოზირება.

ამოცანის პრაქტიკული რეალიზაცია

პრობლემა იყო მონაცემების აქ ქონა რაზეც შევძლებდით მოდელის გაწვრთნა, ამის გადასაჭრელად მოვიძიეთ საიტები რომლის იმფორმაციის გამოყენებასაც შევძლებდით, მოვიძიეთ სამი ონლაინ ყიდვა გაყიდვის ვებ გვერდი myhome.ge, ss.ge და place.ge. შემდეგი პრობლემა გახდა ამ მონაცემების სისწორე და გამოყენებადობა, იყო ბევრი არსწორი მონაცემები შეყვანილი რაც მოდელს უშლიდა სწავლაში ხელს, ამის გასაფილტრად ჩავდეთ ლოგიკა და გავფილტრეთ ტექნიკურად.

დასმული ამოცანის გადასაწყვეტად გადავწყვიტეთ პროგრამირების ენა Python გამოყენება. ასევე გამოვიყენეთ Python ის ცნობილი ბიბლიოთეკა მოცემული პრობლემის გადასაჭრელად სახელად Scrapy.

სისტემა მუშაობს BFS პრინციპით. საიტის დათვალიერებას იწყებს საწყისი გვერდიდან ყოველი ლინკი არის წიბო ამ გრაფში ხოლო გვერდები კი წვეროებია. ყოველი ახალი გვერდის აღმოჩენის შემდეგ გადადის ამ გვერდზე და ამოწმებს ეს გვერდი არის თუ არა იმ ტიპის რომელზეც ჩვენთვის საჭირო ინფორმაციაა განთავსებული. თუ ეს გვერდი ჩვენთვის საჭირო ტიპისაა ამ გვერდიდან ვიღებთ საჭირო მონაცემებს და ვინახავთ მონაცემთა ბაზაში რომელიც ჩვენ შემთხვევაში არის MongoDB. მუშაობას შეწყვეტს მაშინ როცა საიტის ყველა გვერდი იქნება ნანახი.

მოცემულმა მეთოდმა საკმაოდ კარგი შედეგი აჩვენა როგორც სიჩქარის ასევე ინფორმაციის სიმრავლის მხრივ. ss.ge, myhome.ge, place.ge დან ინფორმაციის მოსაგროვებლად მოცემულ პროგრამულ უზრუნველყოფას დაჭირდა 2 დღე და შეგროვდა დაახლოებით 200 000 ჩანაწერი.

მოცემული პროგრამული უზრუნველყოფა გაშვებული იყო ერთ სერვერზე. 3 საიტიდან ინფორმაციის მოგროვება მიმდინარეობდა პარალელურ რეჟიმში ამ ერთ სერვერზე. 2 დღის მანძილზე შეგროვდა 200 000 ჩანაწერი. კონკურენტულ გადაწყვეტებს ვერ ვადარებთ რადგან კონკურენტულ გადაგწყვეტებზე ვერაფერი მოვიძიეთ.

დასკვნა

ჩვენმა მეთოდმა შეძლო მოეგროვებინა მონაცემები რაც შემდგომ წარმატებით იქნა გამოყენებული მანქანური სწავლების მეთოდით პროგნოზირებისთვის. გადაწყვეტაში გასაუმჯობესებელია ის რომ თითო საიტიდან მონაცემები გროვდება მიმდევრობით და ამ პროცესის პარალელიზაციის გზით შესაძლებელია რომ ეს მეთოდი საგრძნობლად აჩქარდეს.

ბიბლიოგრაფია

[1] J. W. Luis Perez, "The Effectiveness of Data Augmentation in Image Classification using Deep Learning," 13 December 2017. [Online]. Available: https://arxiv.org/abs/1712.04621.

[2] I. M. M. H. Debidatta Dwibedi, "Cut, Paste and Learn: Surprisingly Easy Synthesis for Instance Detection," 4 August 2017. [Online]. Available: https://arxiv.org/abs/1708.01642.

[3] K. E. B. A. B. Chaitanya Mitash, "A Self-supervised Learning System for Object Detection using Physics Simulation and Multi-view Pose Estimation," 9 March 2017. [Online]. Available: https://arxiv.org/abs/1703.03347.

[4] P. S. Y. C. W. He Huang, "An Introduction to Image Synthesis with Generative Adversarial Nets," 12 March 2018. [Online]. Available: https://arxiv.org/abs/1803.04469.