KAUNO TECHNOLOGIJOS UNIVERSITETAS
INFORMATIKOS FAKULTETAS
PROGRAMŲ INŽINERIJOS KATEDRA

Justas Šalkevičius

Maitinimo įstaigų reitingo prognozavimo statistinių modelių tyrimas

Magistro darbas

Darbo vadovas
Doc. dr. Tomas Blažauskas

Kaunas, 2015

KAUNO TECHNOLOGIJOS UNIVERSITETAS
INFORMATIKOS FAKULTETAS
PROGRAMŲ INŽINERIJOS KATEDRA

Justas Šalkevičius

Maitinimo įstaigų reitingo prognozavimo statistinių modelių tyrimas

Magistro darbas

Vadovas
Doc. dr. Tomas Blažauskas
2015-05

Atliko
IFM-3/1 gr. stud.
Justas Šalkevičius
2015-05-20

Recenzentas
Prof. dr. Rimantas Butleris
2015-05

Kaunas, 2015

AUTENTIŠKUMO PATVIRTINIMAS

Patvirtinu, kad įteikiamas magistro baigiamasis darbas yra:

1. Atliktas mano paties ir nėra pateiktas kitam kursui šiame ar ankstesniuose semestruose;
2. Nebuvo naudotas kitame Institute/Universitete Lietuvoje ir užsienyje;
3. Nenaudoja šaltinių, kurie nėra nurodyti darbe, ir pateikia visą panaudotos literatūros sąrašą.

Vardas, pavardė

(parašas)

Data

Turinys

Maitinimo įstaigų reitingo prognozavimo statistinių modelių tyrimas

Santrauka

Privačioms ir valstybinėms įmonės kaupiant vis didesnius duomenų kiekius, tuo pačiu aktualesniu darosi jų apdorojimas ir sprendimų priėmimas remiantis žiniomis išgautomis iš šių duomenų. Tokio pobūdžio ekspertinių sistemų kūrimui dažnai pasitelkiami mašininio mokymosi algoritmai dirbantys kompiuterių klasteriuose, kurie dėka produktų, kaip *Cloudera, Amazon E3* ar *Microsoft Azure,* tapo lengvai prieinami platesniam naudotojų ratui.

Šiame darbe yra lyginami regresija pagrįsti mašininio mokymosi modeliai, o jų tyrimams naudojami atviri duomenys iš JAV populiarios verslo įstaigų vertinimo programėlės Yelp. Duomenų modelyje yra pateikti verslo įstaigų sąrašai, jų vertinimai ir įvairios juos apibūdinančios savybės. Mašininio mokymosi algoritmų tyrimui iš šios duomenų bazės buvo atrinkti tik maitinimo įstaigų duomenys ir jų vertinimai.

Šiais duomenimis yra apmokomi regresiniai modeliai, gebantys nustatyt šių įstaigų teikiamų paslaugų (bevielio ryšio prieinamumas, rezervacijų priėmimas ir t.t.) bei savybių (kainos, triukšmo lygio ir t.t.) įtaka reitingams ir pagal jas prognozuoti reitingą naujai. Apmokyto algoritmo modelis tiksliausiai prognozuojantis maitinimo įstaigų reitingus yra panaudojamas ekspertinės sistemos sukūrimui.

Darbe yra analizuojami mašininio mokymosi algoritmai tinkami uždavinio sprendimui ir technologijos jų lygiagretinimui kompiuterių klasteryje. Taip pat suformuojamas, ištiriamas ir algoritmų apmokymui paruošiamas maitinimo įstaigų reitingo duomenų modelis bei juo apmokomi ir įvertinami mašininio mokymosi: tiesinės regresijos, sprendimų medžio ir neuroninių tinklų algoritmai. Galiausiai aptariami rezultatai ir pateikiamos išvados.

Raktiniai žodžiai

Tiesinė regresija, neuroniniai tinklai, sprendimų medis, maitinimo įstaigų reitingai, mašininis mokymasis, regresija

Study of restaurant ratings prediction using regression analysis

Summary

Every year organizations collects even more data, therefore data analysis and data supported decision making is becoming more and more relevant. Smart systems which can process big data usually uses machine learning algorithms and are running in computer clusters. Infrastructure for these kinds of systems can be easily provided to anyone by products like *Cloudera, Amazon E3* or *Microsoft Azure.*

In master thesis a comparison between machine learning regression models are being made based on data provided by popular business review mobile application Yelp. Data model consist of business lists, ratings and their attributes. From this data only information about restaurants and provided services like outdoor seating, reservations, noise levels and etc. are used to train regression models. Later one model which manages to fit testing data most accurately is selected and implemented in smart system for restaurant ratings prediction.

Moreover analysis of machine learning algorithms and parallelization techniques are described in the first chapter. Also data model for restaurant rating prediction is formed and studied to be used to train linear regression, decision tree and neural network models. Finally results and conclusions are provided.

Keywords

Linear regression, neural networks, decision tree, restaurant ratings, machine learning, regression

LENTELIŲ SĄRAŠAS

PAVEIKSLĖLIŲ SĄRAŠAS

TERMINŲ BEI SANTRUMPŲ ŽODYNAS

Mašininis mokymasis (angl. *machine learning*) – dirbtinio intelekto sritis, tirianti apsimokančių sistemų kūrimą.

Klasifikavimas (angl. *clasification*) – mašininio mokymosi uždavinys, kurio tikslas priskirti elementą tam tikrai kategorijai.

Tikslinis kintamasis (angl. *target variable*) – kintamasis, kurio priklausomybę nuo kitų duomenų modelio parametrų nustato regresiniai metodai.

Prognozavimas (angl. *prediction*) – tikslinio kintamojo reikšmės nustatymas naudojant regresinį metodą.

Ištęstinis kintamasis (angl. continuous variable) – kintamasis galintis įgyti reikšmę iš racionalių skaičių aibės.

Neapsimokymas (angl. *underfitting*) – modelis permažai prisitaikęs prie mokymosi duomenų aibės.

Persimokymas (angl. *overfitting*) – modelis per daug prisitaikęs prie mokymosi duomenų aibės.

Duomenų modelio savybė (angl. *data model feature*) – duomenų modelio elemento tipas nuo kurio priklauso prognozuojama tikslinio kintamojo reikšmė.

Sprendimų miškas (angl. *decision tree*) – mašininio mokymosi algoritmas.

Užtikrintumo koeficientas (angl. *coefficient of determination*) – nurodo kaip tiksliai regresijos modelis atitinka duomenis

ĮVADAS

Populiarėjant didžiųjų duomenų apdorojimo sistemoms ir atsirandant vis naujesniems įrankiams skirtiems palengvinti duomenų išgavimą bei analizę, aktualūs tampa mašininio mokymosi algoritmų tyrimai ir taikymai su realias duomenimis. Ekspertinės sistemos paremtos mašininio mokymosi algoritmais gali padėti geriau suprasti verslo procesus, prognozuoti esamų klientų elgesį ar įvertinti riziką.

Šiame darbe yra susitelkiama į mašininio mokymosi regresinius modelius, suteikiančius galimybę prognozuoti skaitines vertes. Tyrimui pasirinkti trys skirtingi regresiniai modeliai, kurie iš esmės skiriasi veikimo idėjomis ir sudarymo modeliu, taip siekiant palyginti platesnio rato algoritmus, o ne tik jų modifikacijas. Tyrimams pasirinkti šie regresiniai mašininio mokymosi algoritmai:

- Tiesinė regresija
- Sprendimų medis
- Neuroninis tinklas

Siekiant darbui suteikti praktinę vertę, pasirinkta, atlikus duomenų modelio sudarymą ir algoritmų tyrimus, sukurti ekspertinę sistemą naudojant modelį geriausiai atitinkantį testavimo duomenų aibę.

Ieškant duomenų bazės, kuri galėtų būti panaudota algoritmų apmokymui, buvo atkreiptas dėmesys į JAV populiarios verslo vertinimo programėlės Yelp siūlomą duomenų aibę. Yelp programėlės kūrėjai akademiniams tyrimams viešai yra pateikę dalį savo duomenų bazės su šimtais tūkstančių duomenų apie verslus, jų įvertinimus ir komentarus.

Peržvelgus jau atliktus tyrimus su šia duomenų baze, nuspręsta tyrimams panaudoti duomenų bazėje įstaigas aprašančius atributus, pvz.: veikimas visą parą, stovėjimo aikštelė klientams ar rezervacijų priėmimas, nes ši dimensija menkai nagrinėta kitų tyrėjų. Kadangi pateiktoje duomenų bazėje aprašomi įvairių sričių verslai, jie turi labai skirtingus atributus, sunku reikšmingai palyginti vaistinės, baro ir parduotuvės teikiamas paslaugas.

Todėl buvo pasirinkta tirti didžiausią aibę turinčią panašius atributus – maitinimo įstaigas, tai kavinės, barai, restoranai ir pan. Jie dalinasi tokioms savybėms, kaip rezervacijos, galimybė sėdėti lauke, alkoholio pasirinkimas ir t.t.

Yelp programėlės autoriai skatina verslo įstaigų savininkus ir lankytojus, suteikti kuo daugiau informacijos apie įstaigą ir teikiamas paslaugas. Tyrimo metu bus nustatyta ar iš tiesų teigiamos savybės ir papildomos paslaugos, kaip rezervacijų priėmimas ar nemokamas internetas lankytojams, teigiamai atsispindi ir įstaigos reitinguose.

Apmokius mašininio mokymosi algoritmus ir vieną iš jų pritaikius ekspertinės sistemos kūrimui. Susidarys galimybė nurodžius įvairias maitinimo vietos savybes (internetas, triukšmo lygis, pageidaujamas apranga, kainų lygis), net neapsilankius įstaigoje prognozuoti jos reitingą.

Ekspertinė sistema su apmokytu modeliu, taip pat leis atrasti kokios teikiamų paslaugų ir savybių kombinacijos suteikia didžiausią reitingo prognozę. Taipogi tirti reitingų kitimo tendenciją, naujiems verslininkams atsižvelgti į labiausiai lankytojų vertinamas paslaugas.

1. ANALITINĖ DALIS

1.1. Analizės tikslas

Analizės dalyje siekiama apžvelgti mašininio mokymosi algoritmus tinkančius maitinimo įstaigų reitingo prognozės uždaviniui spręsti. Taip pat apžvelgiami metodai skirti šių algoritmų naudojimui kompiuterių klasteriuose.

Tyrimo rezultatai bus naudojami maitinimo įstaigų reitingo duomenų modelio sudarymui bei mašininio mokymosi algoritmų parinkimui, kurie bus tiriami.

1.2. Tyrimo sritis, objektas ir problema

Tyrimo objektas – mašininio mokymosi regresijos algoritmai, maitinimo įstaigų reitingo prognozavimui.

Tyrimo problema – siekiant išspręsti maitinimo įstaigų reitingo prognozavimo uždavinį ir tam sukurti ekspertinę sistemą, kyla klausimas, kaip sudaryti duomenų modelį šiai problemai spręsti ir kurį iš mašininio mokymosi algoritmų panaudoti, norint pasiekti kuo tikslesnius rezultatus.

1.3. Tyrimo tikslas ir uždaviniai

Tyrimo tikslas – ekspertinei sistemai parinkti ir apmokinti mašininio mokymosi algoritmą tinkamą maitinimo įstaigų reitingų prognozės uždavinio sprendimui.

Šiam tikslui pasiekti iškelti šie uždaviniai:

- Pasirinkti mašininio mokymosi algoritmus tinkamus reitingo prognozavimui
- Sudaryti ir ištirti maitinimo įstaigų reitingo duomenų modelį
- Įvertinti kiekvieno pasirinkto algoritmo tikslumą
- Pasirinkti algoritmą tinkamiausią ekspertinės sistemos kūrimui

1.4. Reikšmių prognozavimas naudojant statistinius regresinius modelius

Šiame skyriuje yra aptariami egzistuojantys mašininio mokymosi algoritmai skirti duomenų modelio tikslinio kintamojo (angl. *target variable*) prognozavimui.

1.4.1. Mašininio mokymosi algoritmų taikymas duomenų modelių tyrimams

Pagrindinės mašininio mokymosi sistemų taikymo sritys nurodomos kaip [1]:

1. Klasterizavimas – iš aibės elementų turinčių tam tikrų savybių, sudaromos naujos aibės, elementų klasteriai, kurie siejasi tarpusavyje.
2. Klasifikavimas – iš aibės elementų turinčių tam tikrų savybių, parenkama savybė pagal kurią elementai gali būti kategorizuoti.
3. Regresija – iš aibės elementų turinčių tam tikrų savybių, parenkama skaitinė savybė, kuriai yra nustatoma priklausomybė nuo likusių savybių.
4. Rekomendavimas – pagal aibę elementų ir jų savybes - naujam elementui surandami panašiausi į jį.
5. Dažniausios aibės išgavimas – pagal aibę elementų ir jų savybės – naujam elementui surandama aibė, kurioje dažniausiai galima rasti panašų į jį.

Tiriamam maitinimo įstaigų reitingų prognozavimui galima naudoti – klasifikavimą ir regresiją, aprašomus sekančiuose skyreliuose.

1.4.1.1. Mašininio mokymosi klasifikavimo modeliai

Klasifikavimo uždaviniuose sprendžiama problema, kuriai kategorijai priskirti naują elementą, kuris dar nepriklauso jokiai kategorijai, remiantis skaičiavimais padarytais iš apmokymo duomenų aibės [2]. Kiekvienas prognozavimas atliekamas atsižvelgiant į skaičiuojamas elemento savybes.

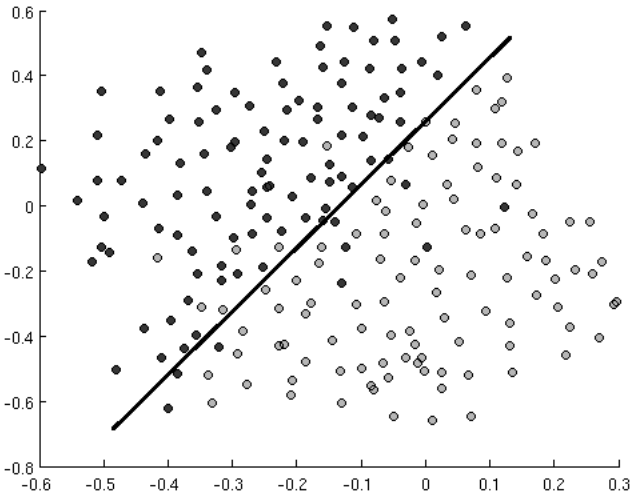
Mašininio mokymosi srityje klasifikavimas priskiriamas, prižiūrimo mokymosi sričiai, t.y. mokymuisi, kai su pateikta duomenų aibe kiekvienam elementui pateikti pilni duomenys - teisingi atsakymai.

Algoritmas, kuris implementuoja klasifikavimo problemos sprendimą, vadinimas klasifikatoriumi, matematine funkcija priskiriančia dėmenį kategorijai.

Klasifikavimo uždaviniui naudojamo duomenų modelio elementų savybių kintamųjų tipai:

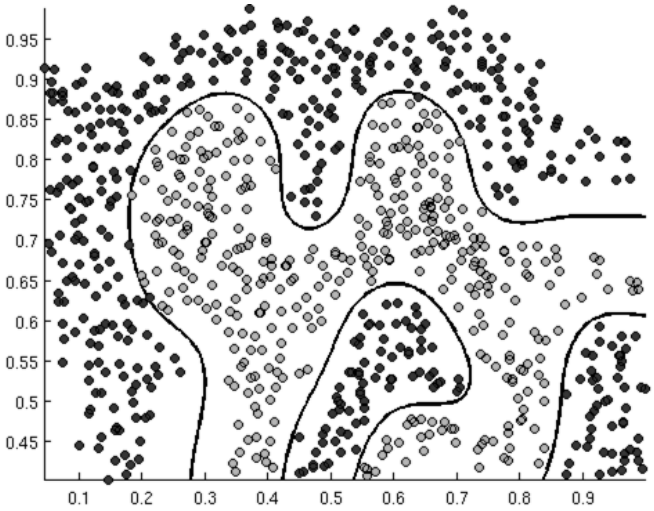
- Kategoriniai – paprastos kategorijos, pvz. kraujo tipas: A, B, O.
- Skaitvardžiai – skaičiuojama aibė, kurios elementai neturi jokios skaitinės įtakos.
- Sveikieji skaičiai – paprastos skaitinės vertės.
- Realieji skaičiai – skaičiai iš realiųjų skaičių aibės.

Tiesinis klasifikavimo modelis yra pats paprasčiausias būdas atskirti duomenų aibę atsižvelgiant į jos elementų savybes, tam yra naudojamas pirmo laipsnio polinomas, kuris esant dviem elemento savybėms yra kreivė, atskirianti dvi plokštumas (1 pav.).



1 pav. Tiesinis klasifikavimas

Dažniausiai tikro pasaulio duomenų modelius sunku atskirti naudojant tiesinį klasifikavimo metodą, ypač esant daugiau nei dviem kategorijoms [3]. Toks klasifikavimo būdas padeda atskirti įvairaus išsidėstymo duomenų aibės elementų sektorius vienas nuo kito (2 pav.).



2 pav. Netiesinis klasifikavimas

1.4.1.2. Mašininio mokymosi regresijos modeliai

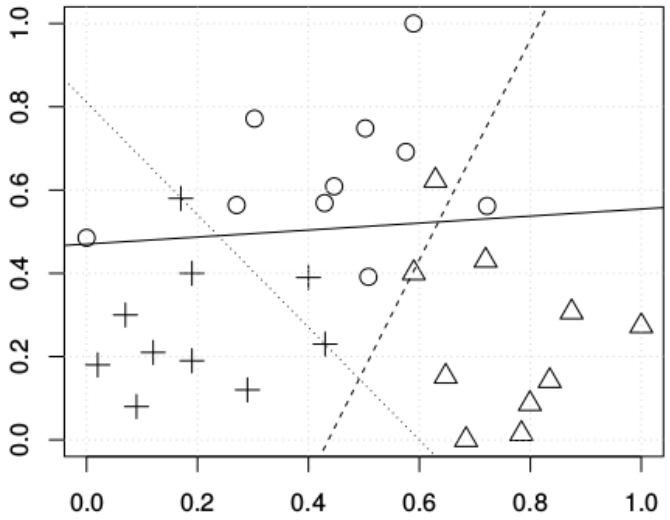
Regresija modeliuoja priklausomybę tarp kintamųjų, iteruodama ir mažindama paklaidą tarp realios ir prognozuojamos reikšmės. Šie regresijos metodai yra pasiskolinti iš statistikos mokslo. Žodis regresija apibrėžia ir uždavinių aibę, ir algoritmų aibę, tačiau bendrai regresiją galima traktuoti kaip procesą.

Paprastai regresiniai modeliai yra taikomi reikšmių prognozavimui, todėl yra tinkami maitinimo įstaigų reitingo prognozės uždavinio sprendimui. Pati paprasčiausia tiesinė regresija traktuoja tikslinio kintamojo reikšmę *y*, kaip savybių *x* kombinaciją. Jei modeliui pateikiamos *x* reikšmės yra be tikslinio kintamojo, modelis tiesiškai pagal daugiklius nustato *y* reikšmę.

1.4.1.3. Klasifikavimo ir regresijos metodų taikymas prognozavimui

Reitingų prognozavimo problemą galima spręsti regresijos arba klasifikavimo metodais (nes egzistuoja maža klasių aibė), šiame skyrelyje aptariami šių metodų taikymo skirtumai.

Klasifikacijos metodai padalina duomenų aibę į klases pagal tikslinį kintamąjį. Paprastai šis kintamasis turi dvi klases: taip arba ne (0 arba 1). Jei yra daugiau nei dvi klasės galima naudoti metodą vienas prieš visus (angl. *one vs. all*), kai parenkama viena klasė, o visos kitos priešingos jai yra traktuojamos kaip antroji klasė (3 pav.).



3 pav. Klasifikavimas esant daugiau nei dviem klasėms

Regresija naudojama kai tikslinis kintamasis yra skaitinis, pavyzdžiui prekės kaina, todėl regresija dažniau taikoma problemoms susijusioms su prognozėmis [4].

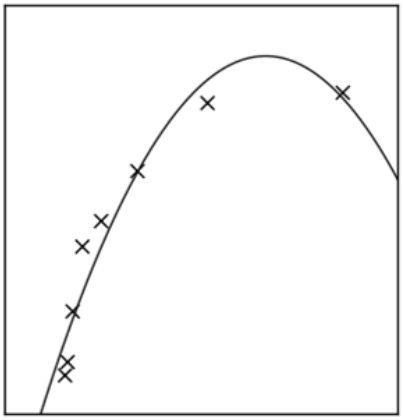
Klasifikacija remiasi duomenų aibės skaidymu pagal jos homogeniškumą. Tarkime turime du kintamuosius žmogaus amžių ir svorį, bei norime nustatyti ar jis valgo greito maisto užkandinėse ar ne. Jeigu apmokymo aibėje 95% žmonių, kuriems iki 30 metų lankosi juose, galime šioje vietoje skaidyti duomenis ir amžius tampa viršutine medžio viršūne, tokiu būdu galime šiuos duomenis pavadinti „grynais“. Klasifikavime duomenų „ne grynumui“, skaičiuojant

pagal duomenų dalį priklausančią aibei, įvertinti dažnai naudojama entropija, nusakanti duomenų aibės homogeniškumą.

Regresija remiasi idėja, jog tikslinis kintamasis neturi klasės, regresijos modelis jam yra pritaikomas naudojant kiekvieną duomenų modelio savybę individualiai. Toliau duomenų aibė yra suskaidoma pagal kiekvieną savybės kintamąjį individualiai, keliuose taškuose. Tada dažniausiai matuojama paklaida tarp prognozuojamos ir realios vertės kiekviename taške ir apskaičiuojama kvadratinių paklaidų suma, galiausiai rekursyviai ieškomi suskaidymo taškai grąžinantys mažiausią paklaidų sumą.

1.4.2. Permokymo ir neapsimokymo problematika mašininiame mokymesi

Apmokant algoritmus svarbu atsižvelgti į jų parametrus ir mokymo aibės dydį. Testuojant algoritmo veikimą – jo atliekamas prognozės, galima stebėti kaip stipriai apmokytas algoritmas yra susiejamas su mokymo duomenų aibe. Tinkamiausiai apsimokęs algoritmas beveik atitinka pateiktos aibės elementus (4 pav.).

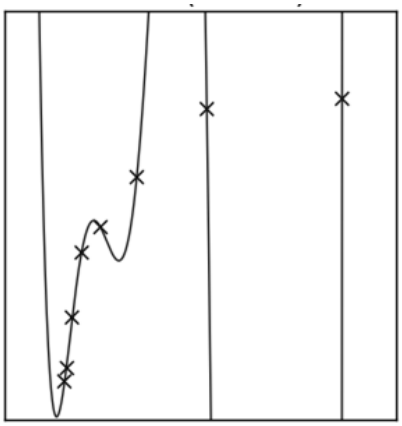


4 pav. Tinkamai apmokyto modelio prisitaikymas

Tačiau dažnai pastebimi du ryškūs nuokrypiai [5]: nepakankamas apsimokymas (angl. *underfitting*) ir persimokimas (angl. *overfitting*). Pastebėjus šiuos reiškinius galima nustatyti, kuriuos algoritmo parametrus reiktų koreguoti siekiant tikslesnių prognozių.

1.4.2.1. Persimokymas (angl. *overfitting*)

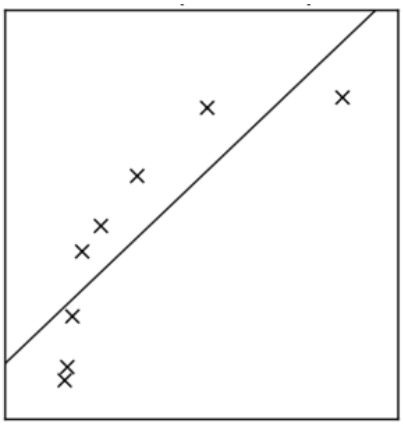
Kai modelis per daug stipriai atitinka duomenų modelio aibę su kuria buvo apmokytas, tai vadinama persimokymu [6]. Vykdant prognozavimą su naujais duomenimis pradeda atsirasti paklaidos priklausančios nuo to kiek nauji elementai atitinka buvusius mokymo aibėje. Tokiu atveju galima koreguoti algoritmo parametrus, pvz. mažinti naudojamo polinomo laipsnį (5 pav.).



5 pav. Modelis per daug prisitaikęs prie apmokymo aibės

1.4.2.2. Neapsimokymas (angl. *underfitting*)

Nepakankamas algoritmo apmokymas (6 pav.) gali įvykti dėl įvairių priežasčių, viena iš dažniausiai pasitaikančių yra per mažas duomenų aibės imties dydis pateikiamas algoritmui.



6 pav. Modelis nepakankamai prisitaikęs prie apmokymo aibės

Tačiau tai gali įtakoti ir įvairūs, algoritmo parametrai, kaip elementų savybėmis taikomi polinomų laipsniai. Pastebėjus nepakankamą apsimokymą, galima bandyti juos pakelti, tačiau tai prailgina apsimokymo laiką.

1.5. Regresiniai mašininio mokymosi algoritmai

1.5.1. Tiesinė regresija

Tiesinė regresija yra naudojama nustatyti empirinę reikšmę. Jo reikšmės skaičiuojamos dauginant apskaičiuotus koeficientus su duomenų modelių elementų savybių reikšmėmis ir ieškant minimalios paklaidos, dažniausiai iteracijos stabdomos sustojus lokaliame minimume [7].

Tiesinei regresijai mašininiame mokymesi dažnai naudojamas greičiausio nusileidimo metodas, kaip pavyzdį paimkime pirmo laipsnio hipotezę (galime įsivaizduoti, kad maitinimo įstaigos reitingas priklauso tik nuo kainos, kurią pažymėkime *x*, norėdami išspręsti šį turime rasti du koeficientus Θ_0 ir Θ_1) (1):

$$h_\Theta(x) = \Theta_0 + \Theta_1 x \text{ (1)}$$

Šią hipotezę įstatykime į paprastą baudos funkciją naudojančią kvadratinę paklaidą (2):

$$J(\Theta_0, \Theta_1) = \frac{1}{2m}\sum_{i=1}^{m}(h_\Theta(x^{(i)}) - y^{(i)})^2 \text{ (2)}$$

Pritaikykime šią baudos funkciją greičiausio nusileidimo metodui (3):

$$kartoti\ iki\ konvergavimo\ \left\{\Theta_j := \Theta_j - \alpha\frac{\partial}{\partial\Theta_j}J(\Theta_0, \Theta_1),\ kai\ j = 1\ ir\ j = 0\right\} \text{ (3)}$$

Pasirinktą baudos funkciją galima supaprastinti iki (4):

$$\frac{\partial}{\partial\Theta_j}J(\Theta) = \frac{1}{m}\sum_{i=1}^{m}(h_\Theta(x^{(i)}) - y^{(i)})x_j^{(i)} \text{ (4)}$$

Gauname algoritmą uždavinio sprendimui (5):

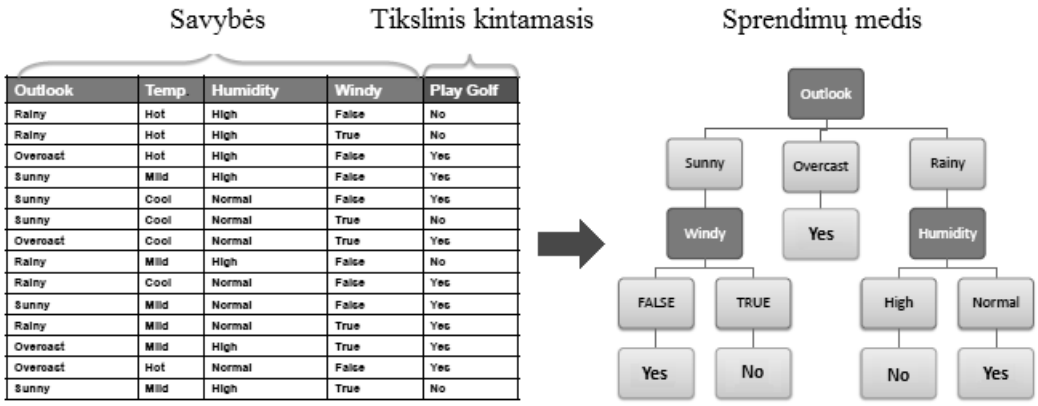
$$kartoti\ iki\ konvergavimo\ \left\{\Theta_j := \Theta_j - \alpha\frac{1}{m}\sum_{i=1}^{m}(h_\Theta(x^{(i)}) - y^{(i)})x_j^{(i)}\right\} \text{ (5)}$$

Tiesinė regresija gali būti parametrizuota šiais parametrais:

1. Polinomu – nusakančiu polinomo laipsnį pagal kurį savybei taikomi apskaičiuoti koeficientai.
2. Mokymosi žingsniu – dydžio tarp iteracijų pokyčiu.
3. Normalizavimo parametru.

1.5.2. Sprendimų medis

Sprendimų medis sukuria klasifikacijos arba regresijos modelius naudodamas medžio struktūrą. Ji padalina duomenis į vis mažesnes ir mažesnes poaibius, tuo pat metu kurdamas susijusius sprendimų medžius (7 pav.). Galutinis rezultatas yra medis su sprendimų viršūnėmis ir lapų viršūnėmis. Sprendimų viršūnės (pvz. porcijos kiekis) turi dvi arba daugiau šakų (pvz. maža, vidutinė, didelė), tuo tarpu lapų viršūnės (pvz. 1 žvaigždutės reitingas) reprezentuoja klasifikacija arba priimamą sprendimą.



Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

7 pav. Sprendimų medžio formavimas iš duomenų aibės

Aukščiausia sprendimų viršūnė medyje atitinka geriausią klasifikatorių ir vadinama šaknine viršūne. Sprendimų medžiai gali apdoroti ir kategorinius, ir skaitinius tikslinius kintamuosius.

Sprendimų medžių formavimui dažnai naudojamos entropijos ir informacijos pasipildymo sąvokos. Medis yra kuriamas nuo viršaus į apačią pradedant nuo šakninės viršūnės, skaidant duomenis į poaibius, kurie turi panašias reikšmes (yra homogeniški) . Homogeniškumas apskaičiuojamas panaudojant entropiją. Jei aibė visiškai homogeniška, entropija lygi nuliui, o jeigu lygiai padalinta, ji lygi vienetui [8].

Entropija vienai savybei yra apskaičiuojama formule (1):

$$E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i \quad (1)$$

$$kur\ i\ savybės\ kategorija, c\ savybės\ kategorijų\ kiekis$$

Entropija kelioms savybėms yra skaičiuojama padauginant iš elemento buvimo poaibyje tikimybės (2):

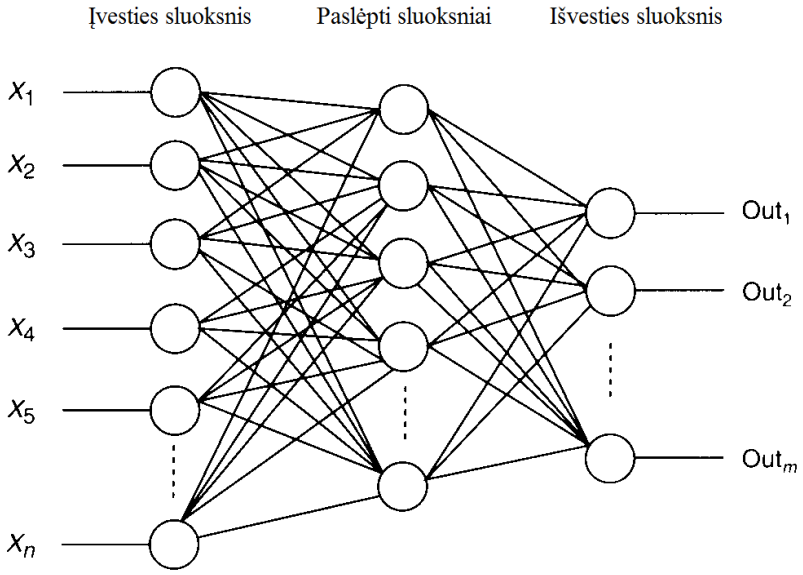
$$E(T, X) = \sum_{c \in X} P(c) E(c) \quad (2)$$

Informacijos pasipildymas remiasi entropijos sumažėjimu, po aibės išskaidymo pagal tam tikrą savybę (3). Konstruojant sprendimų medį pagrindinis tikslas surasti savybes, kurios suteikia didžiausią informacijos pasipildymą (homogeniškiausias šakas).

$$Gain(S, A) = E(S) - \sum_{v \in Reikšmėms(A)} \frac{|S_v|}{|S|} E(S_v) \quad (3)$$

1.5.3. Neuroniniai tinklai

Neuroninį tinklą sudaro dirbtiniai neuronai, dar vadinami viršūnėmis. Šios viršūnės yra sujungtos tarpusavyje ir šioms jungtims yra priskiriama stiprumo vertė yra lygi jų poveikio stiprumui: slopinantis (iki -1) ir skatinantis (iki +1). Jei ši vertė yra didelė, tai simbolizuoja stiprų ryšį tarp viršūnių [9]. Taip pat kiekvienoje viršūnėje yra sugeneruojama perkėlimo funkcija ir jos yra sugrupuojamos į įvesties, paslėptus ir išvesties sluoksnius (8 pav.).



8 pav. Neuroninių tinklų sluoksniais

Įvesties viršūnės priima informaciją išreikštą skaitiniu pavidalu. Šioms įvestims tada pritaikomos aktyvavimo reikšmės ir informacija išnešiojama per visą tinklą. Pagal kraštinių svorius, slopinimo arba skatinimo daugiklius, ir perkėlimo funkcijas bei aktyvavimo reikšmes, informacija perduodama iš viršūnės į viršūnę.

Kiekviena viršūnė sumuoja priimamas aktyvavimo reikšmes ir modifikuoją ją pagal savo perkėlimo funkciją. Toks procesas vyksta per visus paslėptus sluoksnius iki tol kol pasiekiamas išvesties sluoksnis. Išvesties sluoksnio viršūnės galiausiai paverčia informaciją į galutinį rezultatą.

1.6. Mašininio mokymosi algoritmų pritaikymas veikimui kompiuterių klasteryje

Koeficientų reikalingų duomenų aibės klasifikavimui nustatymas dėl paprastai reikalingų bent kelių šimtų iteracijų ir kelių šimtų tūkstančių elementų analizės užtrunka ilgą laiko tarpą, todėl paprastai yra atliekamas kompiuterių klasteryje.

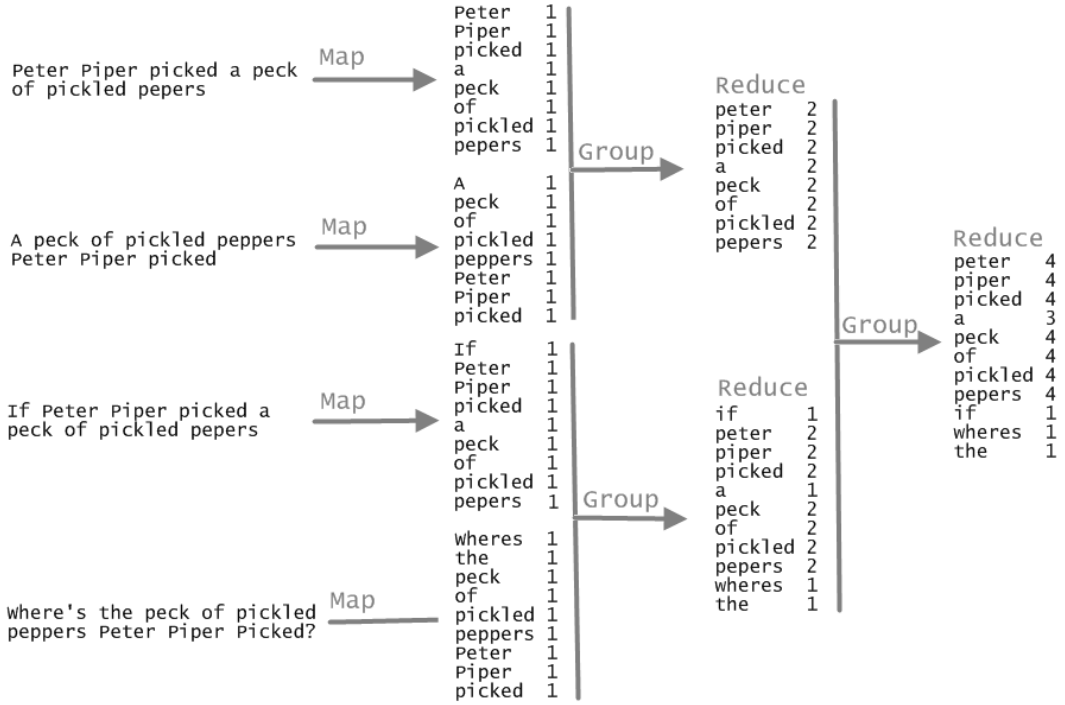
Apdorojant didelius duomenų kiekius paskirstytose sistemose ir norint optimizuoti, tokių sistemų našumą dažnai naudojamas MapReduce modelis, kurio esmė išskirstyti algoritmą į dvi procedūras, kurios galėtų būti pasiskirstytos.

1.6.1. MapReduce modelis duomenų analizei kompiuterių klasteryje

MapReduce tai programavimo modelio implementacija skirta apdoroti ir generuoti dideles duomenų aibes naudojant išlygiagrentintus algoritmus kompiuterių klasteryje.

Šis modelis yra sudarytas iš dviejų procedūrų (9 pav.):

- Map – skirtos duomenų filtravimui ir rūšiavimui;
- Reduce – skirtos rezultatų apibendrinimui;



9 pav. MapReduce algoritmo eiga

MapReduce modeliu grįstos sistemos organizuoja serverių veiklą klasteryje. jų komunikavimą tarpusavyje, duomenų perdavimą ir klaidų valdymą.

Modelis remiasi tokių pačių pavadinimų funkcijomis sutinkamomis funkcinėse kalbose, nors šių funkcijų paskirtis MapReduce sistemose yra visai kita. Tikroji MapReduce sistemų nauda nėra *map* ir *reduce* funkcijų naudojimas, tačiau lengvas implementavimas klasteriuose ir klaidų tolerancija. Todėl algoritmo naudojimas MapReduce sistemoje tik su viena viršūne, nebus greitesnis nei tradicinis įgyvendinimas. Modelio nauda pasireiškia tik naudojant optimizuotą maišymo operaciją, kuri sumažina tinklo apkrovimą ir klaidų valdymo savybę.

Viena iš populiariausių šio modelio atviro kodo implementacija yra Apache Hadoop.

1.6.2. Hadoop karkasas duomenų apdorojimui kompiuterių klasteryje

Apache Hadoop karkasas yra skirtas didelių duomenų aibių saugojimui ir apdorojimui kompiuterių klasteryje. Šį karkasą sudaro:

- Hadoop Common – pagrindinės bibliotekos reikalingos kitoms sistemos funkcijos įgyvendinti;
- Hadoop Distributed File System (HDFS) – paskirstyta failų sistema skirta saugoti duomenų failu tarp klasterio kompiuterių.
- Hadoop YARN – resursų valdymo sistema skirta skaičiavimų valdymui klasteryje.
- Hadoop MapReduce – modelis skirtas didelių duomenų kiekių apdorojimui.

Apart HDFS, YARN ir MapReduce modulių, Apache Hadoop platforma taip pat yra pagrindas daugeliui kitų sistemų skirtų darbui su dideliais duomenų kiekiais, kaip Apache Pig ir Apache Hive skirtų SQL užklausų rašymui HDFS sistemose, Apache HBase – NoSQL duomenų bazė ir Apache Mahout – mašininio mokymosi algoritmų naudojimui.

1.7. Analizės išvados

Atlikus mašininio mokymosi metodų analizę padarytos šios išvados:

1. Maitinimo įstaigų reitingo prognozavimui labiau tinkami regresiniai metodai, nes nepaisant mažos klasių aibės, tikslinis kintamasis yra skaitinė vertė, kurių prognozei taikomi regresiniai metodai.
2. Maitinimo įstaigų reitingo prognozavimo duomenų modelio savybės tipo elementų parinkimas turės didelę įtaką modelio tikslumui.
3. Tiesinės regresijos algoritmas leis nustatyti sudaryto duomenų modelio savybių svorių įtaką įstaigos reitingui (tiesinę priklausomybę).
4. Sprendimo medžio ir neuroninių tinklų algoritmai suteiks galimybę įvertinti netiesinę savybių įtaką maitinimo įstaigos reitingui.
5. Naudojantis MapReduce metodo pagrindu mašininio mokymosi algoritmų apmokymą greičiau galima atlikti kompiuterių klasteryje.

2. PROJEKTINĖ DALIS

2.1. Duomenų apžvalga

2.1.1. Yelp programėlės teikiama duomenų aibė

JAV bendrovė "Yelp", kuri kurią programėlė vietinių verslų apžvalgai, akademiniam naudojimui yra pateikusi dalį savo duomenų iš Phoenix, Las Vegas, Madison, Waterloo ir Edinburgh miestų [10], kuriuose aprašoma informacija apie:

- 42,153 verslo vienetų
- 320,002 verslo savybių
- 31,617 prisiregistravimų
- 252,898 vartotojų
- 403,210 patarimų
- 1,125,458 apžvalgų

Yelp programėlės kūrėjai ragina vartotojus ir įstaigų savininkus užpildyti kuo išsamesnį aprašymą apie teikiamas paslaugas, pvz.:

- Prekiaujamo alkoholio tipas;
- Ar turi DJ;
- Ar grojama foninė muzika;
- Ar turi muzikos mašiną (angl. Jukebox);
- Ar grojama gyva muzika;
- Ar rengiamas karaoke vakaras;
- Ar priimamos kreditinės kortelės;
- Kainos ruožas (nuo 1 iki 5);
- Ar yra lauko "terasa";

Pagal šiuos duomenis yra kuriama ekspertinė sistema bei apmokomi tiriami mašininio mokymosi algoritmai, siekiant nustatyti galimą įstaigos reitingą, pagal jos teikiamas paslaugas ir parametrus.

2.1.2. Yelp programėlės duomenų aibės formatas

Visi duomenys pateikiami 1.3GB dydžio faile JSON formatu, su ne visada teisingu formatavimu ir eiliškumu. Tiriamus verslo vienetus aprašo šis formatas (10 pav.):

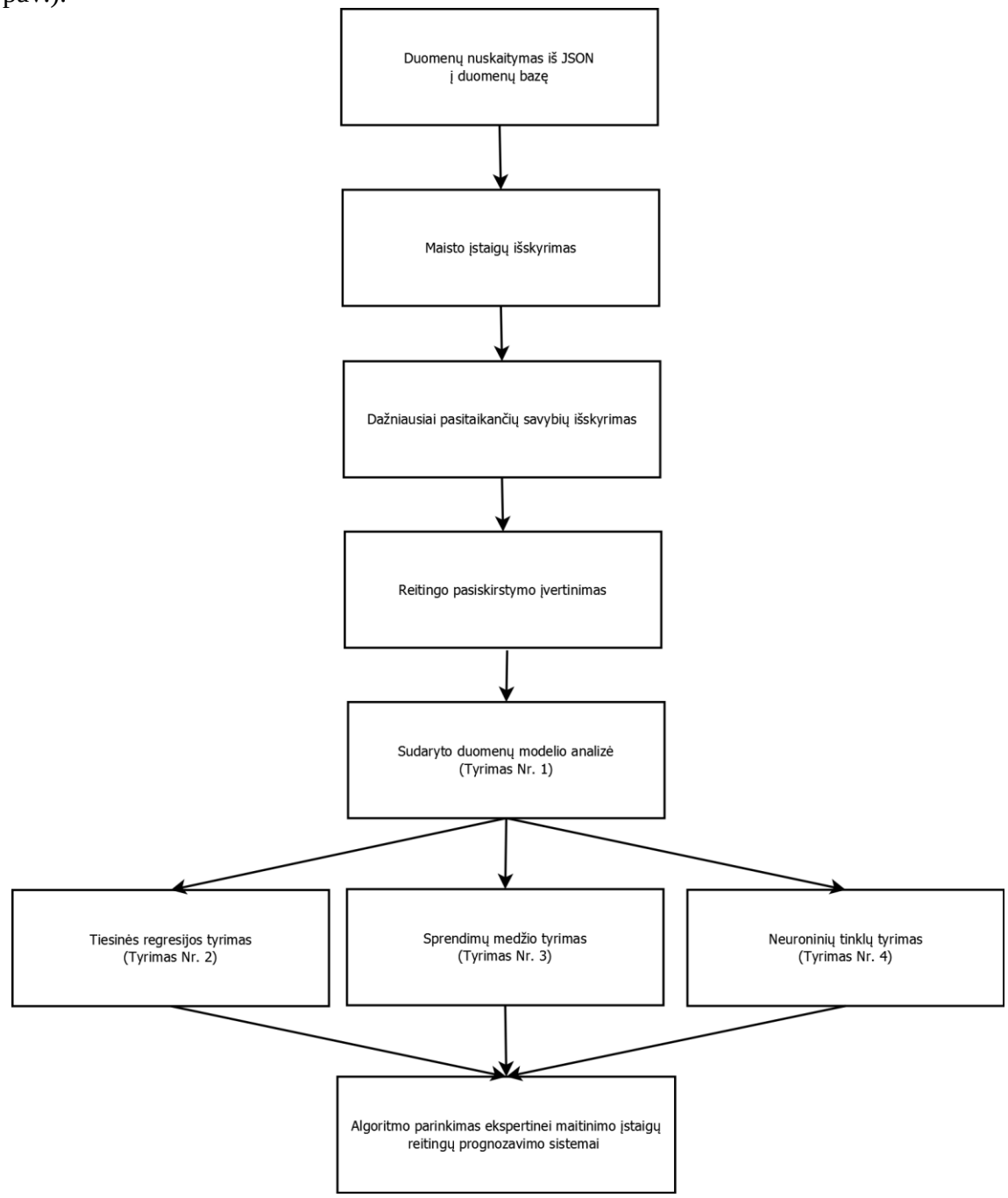
business

```
{
    'type': 'business',
    'business_id': (encrypted business id),
    'name': (business name),
    'neighborhoods': [(hood names)],
    'full_address': (localized address),
    'city': (city),
    'state': (state),
    'latitude': latitude,
    'longitude': longitude,
    'stars': (star rating, rounded to half-stars),
    'review_count': review count,
    'categories': [(localized category names)]
    'open': True / False (corresponds to closed, not business hours),
    'hours': {
        (day_of_week): {
            'open': (HH:MM),
            'close': (HH:MM)
        },
        ...
    },
    'attributes': {
        (attribute_name): (attribute_value),
        ...
    },
}
```

10 pav. Yelp pateikiamų duomenų apie verslo vienetus formatas JSON

2.2. Tyrimo planas

Maitinimo įstaigų reitingo prognozavimo tyrimas suskirstytas į šiuos smulkesnius etapus (11 pav.):



11 pav. Tyrimo plano schema

- **Duomenų nuskaitymas iš JSON į duomenų bazę** – dėl didelio JSON duomenų failo ir jame esančių perteklinių duomenų, tolesniam jų apdorojimui ir analizei efektyviau yra nuskaitytus duomenis išsaugoti duomenų bazėje;
- **Maitinimo įstaigų išskyrimas** – atrenkamos tik tos įstaigos, kurių kategorija yra susijusi su maitinimu;
- **Dažniausiai pasitaikančių savybių išskyrimas** – pateiktuose duomenyse, ne visos įstaigos turi pilnus aprašymus, šiame etape išskiriamos įstaigos turinčios visus tiriamus parametrus;
- **Reitingo pasiskirstymo įvertinimas** – peržiūrimas maitinimo įstaigų reitingų pasiskirstymas ir atliekamas jo papildymas;
- **Sudaryto duomenų modelio analizė (Tyrimas Nr. 1)** – įvertinama reitingo tiesinė priklausomybė nuo jį apibūdinančių duomenų modelio savybių;
- **Tiesinės regresijos tyrimas (Tyrimas Nr. 2)** – apmokomas ir įvertinamas tiesinės regresijos algoritmas;
- **Sprendimų medžio tyrimas (Tyrimas Nr. 3)** – apmokomas ir įvertinamas sprendimų medžio modelis;
- **Neuroninių tinklų tyrimas (Tyrimas Nr. 4)** – apmokomas ir įvertinamas neuroninis tinklas;
- **Algoritmo parinkimas ekspertinei maitinimo įstaigų reitingo prognozavimo sistemai –** įvertinami tyrimų nr. 2, 3 ir 4 rezultatai, bei parenkamas algoritmas

2.2.1. Maitinimo įstaigų išskyrimas

Iš pradinės duomenų aibės sudarytos iš vairių verslo sričių, atrinktos įmonėms, kurioms duomenų aibėje priskirta maitinimo įstaigoms būdinga kategorija:

- Restoranai;
- Barai;
- Sporto barai;
- Smuklės;
- Kavinės;
- Užeigos;
- Vyno barai;
- Koktelių barai;
- Valgyklos;
- Alaus barai;

2.2.2. Dažniausiai pasitaikančių savybių išskyrimas

Siekiant užtikrinti duomenų pilnavertiškumą, nustatytos dažniausiai užpildytos savybės apie maitinimo įstaigas (1 lentelė).

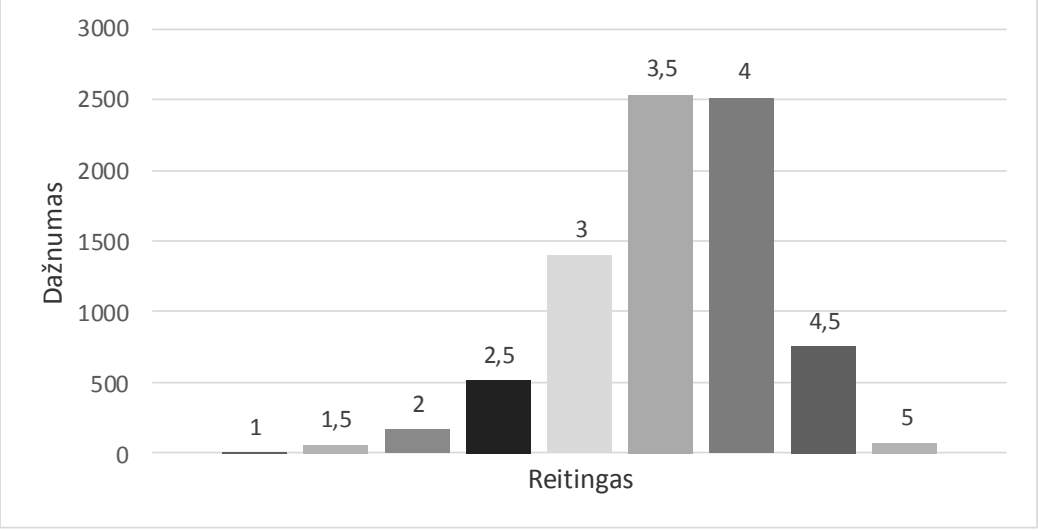
1 lentelė Maitinimo įstaigų savybių užpildymas Yelp duomenų modelyje

Savybė	Užpildymas (%)
Kaina	93
Galimybė sėdėti lauke	83.6
Alkoholio pasirinkimas	74.6
Pageidaujama apranga	74
Rezervacijos galimybė	71.6
Triukšmo lygis	66.8
Bevielio ryšio prieinamumas	57

Tolesniam tyrimui buvo paliktos tik tie duomenų aibės elementai, kurie turėjo informaciją apie visas atrinktas savybes, tokiu būdu užtikrinant pilnavertišką algoritmų apsimokymą ir išsprendžiant trūkstamų reikšmių interpretavimo problemą.

2.2.3. Reitingo pasiskirstymo įvertinimas

Apžvelgiant reitingų pasiskirstymą Yelp programėlės duomenų bazėje, pastebime jog yra gana nedaug labai mažų reitingų bei labai didelių reitingų elementų (12 pav.).



12 pav. Pradinis reitingų pasiskirstymas duomenų modelyje

Todėl siekiant pabrėžti blogas ir geras savybes, esami reitingai buvo perskaičiuoti 3 žvaigždučių sistemai. Naujas paskirstymas buvo atliktas sekančiu metodu:

- Žemas – 1 žvaigždutės reitingas (jam priskirti buvę reitingai nuo 1 iki 2.5 žvaigždučių);
- Vidutinis – 2 žvaigždučių reitingas (jam priskirti buvę reitingai nuo 3 iki 4 žvaigždučių);
- Aukštas – 3 žvaigždučių reitingas (jam priskirti buvę reitingai nuo 4.5 iki 5 žvaigždučių);

3. TYRIMO IR EKSPEREMENTINĖ DALIS

3.1. Atliekamo tyrimo metodologija

Šiame darbe yra atliekami maitinimo įstaigų reitingo prognozės duomenų modelio sudarymo ir mašininio mokymosi regresijos algoritmų tyrimai. Algoritmai yra apmokomi naudojant duomenis iš akademinės Yelp duomenų bazės, joje išfiltravus duomenis apie maitinimo įstaigas.

Tyrimams sudarytą maitinimo įstaigų duomenų modelį, kurio paruošimas aprašytas 2 skyriuje, sudaro sekantys parametrai:

2 lentelė Maitinimo įstaigų duomenų modelio parametrai

Pavadinimas	**Tipas**	**Reikšmės**
Kainų lygis (angl. *price range)*	Savybė	Įvertinimas 1 – 5; 1 – žemiausias lygis (pigiausia) 5 – aukščiausias lygis (brangiausia)
Galimybė sėdėti lauke (angl. *outdoor seating*)	Savybė	Yra; Nėra;
Alkoholis (angl. *alcohol*)	Savybė	Nėra (angl. *none*); Pilnas pasirinkimas (angl. *full bar*); Alus ir vynas (angl. *beer and wine*);
Pageidaujama apranga (angl. *attire*)	Savybė	Įprasta (angl. *casual*); Prašmatni (angl. *dressy*); Formali (angl. *formal*);
Galimybė rezervuotis staliuką (angl. *reservations*)	Savybė	Yra; Nėra;
Triukšmo lygis (angl. *noise level*)	Savybė	Tylus (angl. *quiet*);

		Vidutinis (angl. *average*); Garsus (angl. *loud*); Labai garsus (angl. *very laud*);
Bevielio ryšio prieinamumas (angl. *Wi-Fi*)	Savybė	Nėra (angl. *none*); Nemokamas (angl. *free*) Mokamas (angl. *paid*)
Reitingas	Prognozuojama reikšmė	Įvertinimas 1 – 3; 1 – žemiausias lygis (vertinama neigiamai) 3 – aukščiausias lygis (vertinama teigiamai)

Mašininio mokymosi algoritmų tyrimai yra vykdomi kompiuterių klasteryje, juos sudaro dvi dalys:

1. Algoritmo modelio apmokymas, naudojant 70 proc. sudaryto duomenų modelio elementų, parinktų atsitiktinai;
2. Apmokyto algoritmo modelio įvertinimas, naudojant likusius 30 proc. sudaryto duomenų modelio elementų, parinktų atsitiktinai, kurie nebuvo naudoti apmokant algoritmą;

Apmokyti mašininio mokymosi algoritmų modeliai yra įvertinami pagal jų paklaidas, apskaičiuotas naudojant testavimo duomenų aibę (antras žingsnis), kuri nebuvo naudota apmokant algoritmus.

Algoritmų vertinimui naudojamos šios reikšmės:

- Vidutinė absoliutinė paklaida (angl. *mean absolute error*) – dydis nusakantis, kiek skiriasi prognozuojamos reikšmės nuo tikrų reikšmių. Turi tokį pat matavimo vienetą kaip ir prognozuojama reikšmė, todėl gali būti lyginama tik tarp to paties duomenų modelio.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i|$$

kur f_i prognozė ir y_i tikroji reikšmė

- Šaknis iš vidutinės kvadratinės paklaidos (angl. root *mean squared error*) – turi tokį pat matavimo vienetą kaip ir prognozuojama reikšmė, todėl gali būti lyginama tik tarp to paties duomenų modelio.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(f_i - y_i)^2}{n}}$$

kur f_i prognozė ir y_i tikroji reikšmė

- Reliatyvi absoliutinė paklaida (angl. *relative absolute error*) – gali būti lyginama ir tarp modelių turinčių skirtingus matavimo vienetus.

$$RAE = \frac{\sum_{i=1}^{n}|f_i - y_i|}{\sum_{i=1}^{n}|\overline{y} - y_i|}$$

kur f_i prognozė, y_i tikroji reikšmė ir $\overline{y}$ yra y_i reikšmių vidurkis

- Reliatyvi kvadratinė paklaida (angl. *relative squared error*) – gali būti lyginama ir tarp modelių turinčių skirtingus matavimo vienetus.

$$RSE = \frac{\sum_{i=1}^{n}(f_i - y_i)^2}{\sum_{i=1}^{n}(\overline{y} - y_i)^2}$$

kur f_i prognozė, y_i tikroji reikšmė ir $\overline{y}$ yra y_i reikšmių vidurkis

- Užtikrintumo koeficientas (angl. *coefficient of determination*) – nurodo kaip tiksliai regresijos modelis atitinka duomenis

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

$$SSR = \sum_i (f_i - \bar{y})^2$$

$$SST = \sum_i (y_i - \bar{y})^2$$

$$SSE = \sum (y_i - f_i)^2$$

kur f_i prognozė, y_i tikroji reikšmė ir $\overline{y}$ yra y_i reikšmių vidurkis

Šiame darbe yra nutarta vykdyti tokius tyrimus:

1. Maitinimo įstaigų duomenų modelio elementų reikšmių (bevielio ryšio prieinamumas, triukšmo lygis ir t.t.) įtakos, prognozuojamai reikšmei – reitingui, tyrimas;
2. Mašininio mokymosi tiesinės regresijos algoritmo tyrimas naudojant sudarytą duomenų modelį ir prognozuojant maitinimo įstaigos reitingą pagal jos savybes ir teikiamas paslaugas;
3. Sprendimų medžio algoritmo tyrimas naudojant sudarytą duomenų modelį ir prognozuojant maitinimo įstaigos reitingą pagal jos savybes ir teikiamas paslaugas;
4. Neuroninių tinklų algoritmo tyrimas naudojant sudarytą duomenų modelį ir prognozuojant maitinimo įstaigos reitingą pagal jos savybes ir teikiamas paslaugas;

3.2. Pasirinktų tyrimo metodų paskirtis

3 lentelė Pasirinktų tyrimo metodų aprašas ir paskirtis

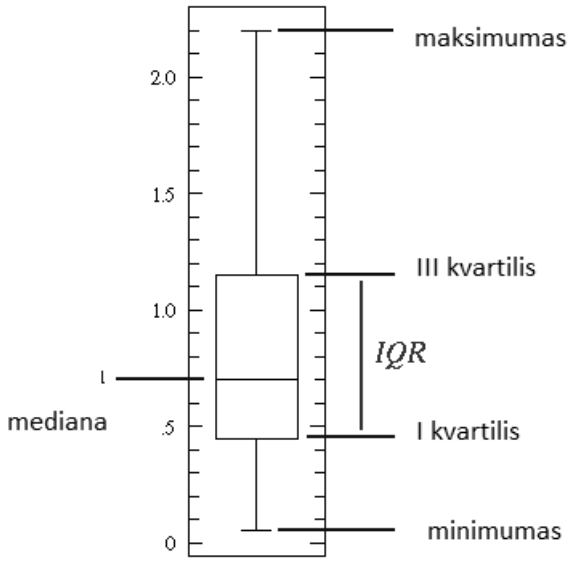
Tyrimo numeris	Paskirtis
1	Tyrimas atliekamas siekiant išsiaiškinti ar egzistuoja koreliacija tarp duomenų modelio savybių ir tikslinio kintamojo (prognozuojamos reikšmės). Ištyrus sudarytą maitinimo įstaigų reitingo duomenų modelio parametrų įtaką reitingui galima nusakyti ar egzistuoja tiesinė reitingo priklausomybė nuo atskirų elementų.
2	Apmokyti ir įvertinti mašininio mokymosi tiesinės regresijos metodo tikslumą prognozuojant maitinimo įstaigų reitingus. Nustatyti duomenų modelio savybių įtakos reitingui koeficientus, naudojimus jo prognozavimui.
3	Apmokyti ir įvertinti mašininio mokymosi sprendimo medžio metodo tikslumą prognozuojant maitinimo įstaigų reitingus. Šis metodas sudarys netiesinės parametrų priklausomybės modelį.
4	Apmokyti ir įvertinti mašininio mokymosi neuroninių tinklų metodo tikslumą prognozuojant maitinimo įstaigų reitingus. Šis metodas sudarys netiesinės parametrų priklausomybės modelį pagal neuroninio tinklo grafą.

3.3. Tyrimo rezultatai

3.3.1. Tyrimas Nr. 1

Tyrimo metu siekiama išsiaiškinti koreliaciją tarp maitinimo įstaigų reitingo duomenų modelio savybių ir tikslinio kintamojo (reitingo). Šis tyrimas padės įvertinti prognozuojamus tiesinės regresijos metodo rezultatus.

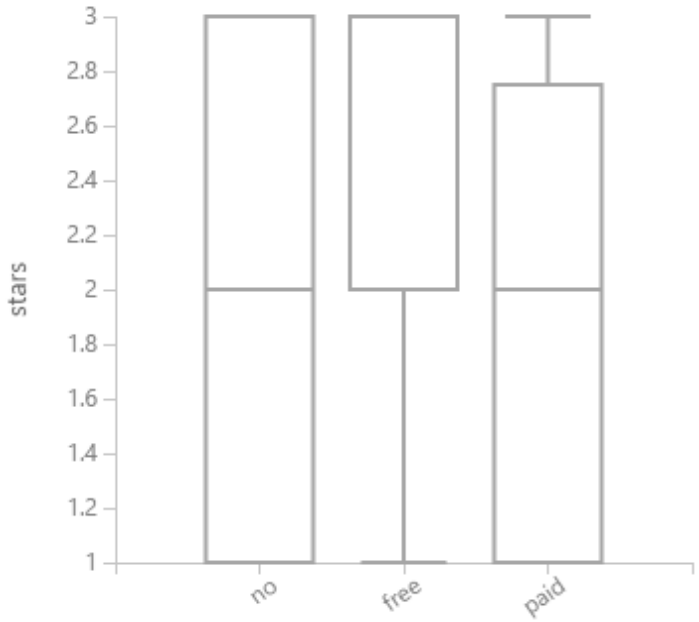
Duomenų savybių priklausomybės atvaizdavimui pasirinktos dėžutės tipo diagramos (angl. *box plot*) (13 pav.), dažnai naudojamos didelių duomenų kiekio analizei, jas sudaro:



13 pav. Dėžutės tipo diagrama

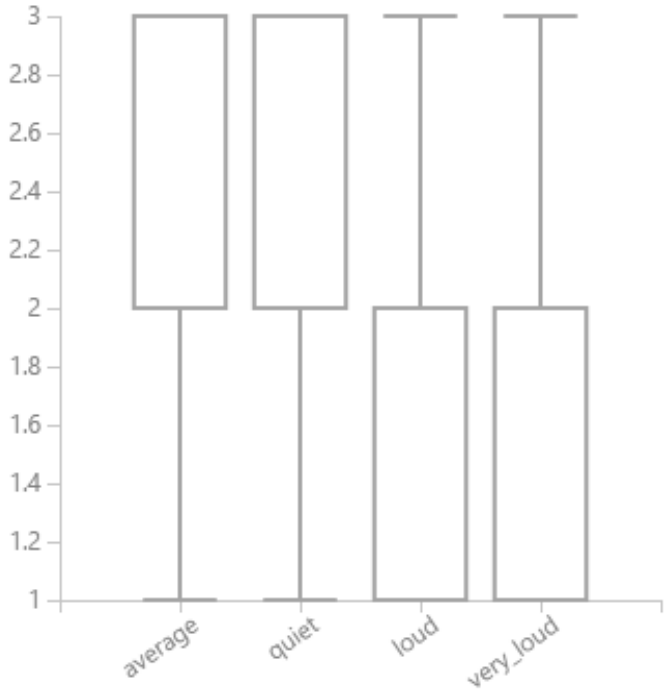
- Minimumas, maksimumas ir mediana
- Pirmas kvartilis – atkerta apatinius 25 proc. duomenų
- Trečias kvartilis – atkerta viršutinius 25 proc. duomenų

Ištyrę duomenų modelį pagal bevielio interneto prieinamumą lankytojams (14 pav.), galime pasakyti, kad mokamas bevielis ryšys maitinimo įstaigoje daro neigiamą įtaką jos reitingui, palyginus su nemokamu, tokio rezultato turėtume tikėtis ir tiesinėje regresijoje.



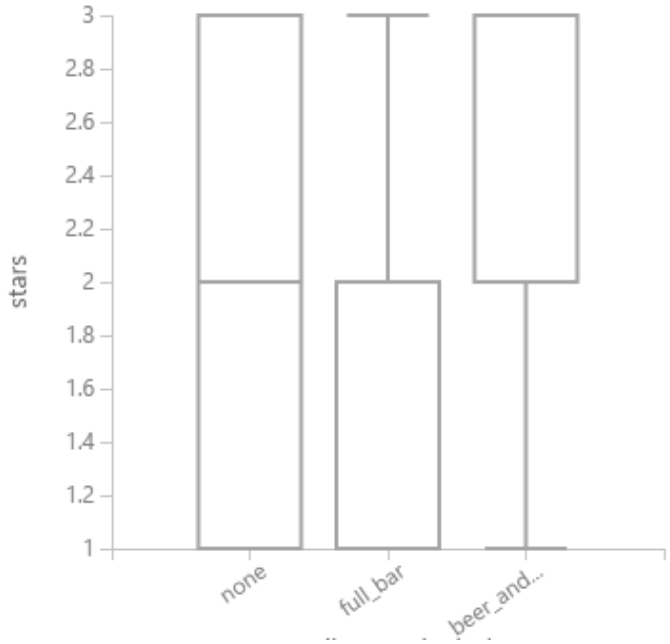
14 pav. Bevielio ryšio prieinamumo savybės įtaka reitingui

Ištyrę duomenų modelį pagal garso lygį (15 pav.), galime pasakyti, jog egzistuoja akivaizdi tiesinė reitingo priklausomybė nuo šio parametro ir lankytojai labiau teigiamai vertina tylias ar vidutiniškai triukšmingas maitinimo įstaigas, nei garsias arba labai garsias. Todėl tiesinės regresijos metodas turėtų nustatyti atitinkamai labiau teigiamus arba neigiamus koeficientus.



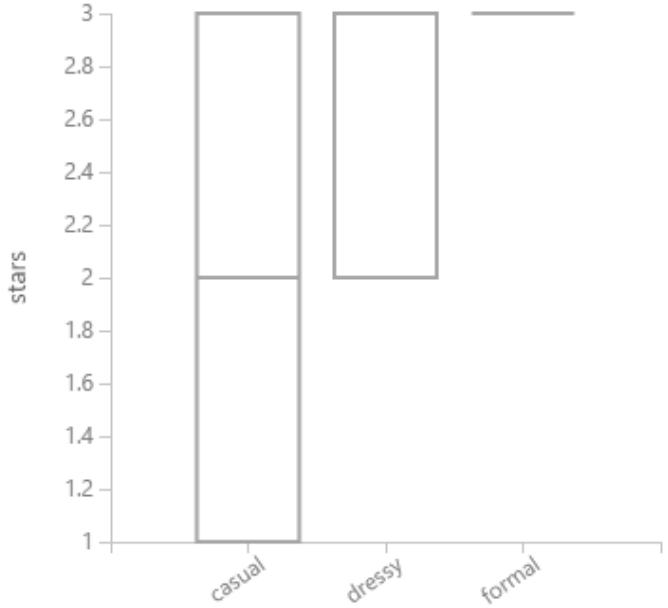
15 pav. Triukšmo lygio savybės įtaka reitingui

Ištyrę duomenų modelį pagal alkoholio pasirinkimą (16 pav.), galime pasakyti, jog didžiausią teigiamą įtaką maitinimo įstaigos reitingui tiesinėje regresijoje turėtų daryti specializavimasis alumi arba vynu.



16 pav. Alkoholio pasirinkimo įtaka reitingui

Ištyrę duomenų modelį pagal pageidaujamą aprangą (17 pav.), galime pasakyti, jog teisinė regresija aukštesnį reitingą turėtų priskirti maitinimo įstaigoms, kurios laukia pasipuošusių ar formaliai apsirengusių klientų.



17 pav. Pageidaujamos aprangos įtaka reitingui

3.3.2. Tyrimas Nr. 2

Tyrimui pasirinktas mažiausių kvadratų tiesinė regresija (angl. *least squares linear regression*) yra vienas iš populiariausių prognozinės analizės (angl. *predictive analytics*) metodų. Šis metodas priima, jog egzistuoja stipri tiesinė priklausomybė tarp duomenų modelio savybės elementų ir prognozuojamojo kintamojo.

Šis metodas naudoja įprastinę mažiausių kvadratų (angl. *ordinary least squares*) baudos funkciją, kuri paklaidą apskaičiuoja kaip kvadratų tarp prognozuojamos ir realios reikšmės sumą, tokiu būdu pritaikant modelį minimizuojama kvadratinė paklaida.

Šiam modeliui galima parinkti šį parametrą:

- **Reguliarizacijos svoris** – naudojimas permokymo išvengimui, tyrimo metu šiam parametrui buvo priskirta reikšmė 0.001

Atlikus tiesinės regresijos modelio apmokymą su paruoštu duomenų modeliu buvo gauti parametrų svoriniai įverčiai (4 lentelė).

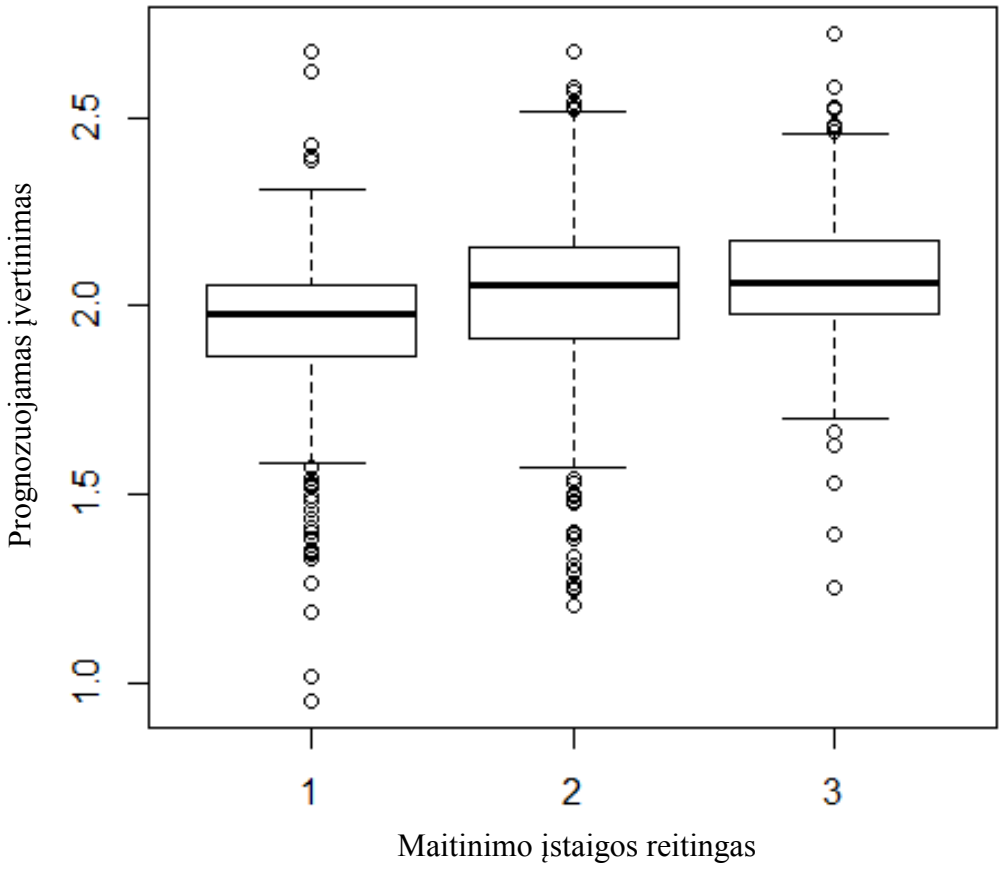
Galima pastebėti, jog gauti įverčiai atitinka ir sukonkretizuoja įvairių savybių įtaką reitingui, nustatytų Tyrimas Nr. 1 metu. Reitingą smarkiai augina specializavimasis alumi arba vynu, mokamas bevielis ryšys reitingą mažina, o nemokamas kelia, didelės kainos turi neigiamą įtaką reitingui, prabangūs restoranai laukiantys pasipuošusių lankytojų yra vertinami geriau, klientai geriau vertina tylias vietas, nei triukšmingas ir t.t.

Todėl akivaizdu, kad šie svoriai atspindi teigiamas ar neigiamas maitinimo įstaigų savybes, kurių galime tikėtis ir realiose pasaulio situacijose, todėl suformuotą reitingo prognozės modelį galima naudoti ekspertinės sistemos kūrimui.

4 lentelė Tiesinės regresijos svoriniai įverčiai

	Reikšmė	**Svoris**
Alkoholio pasirinkimas	Alus ir vynas	0.42839
	Pilnas pasirinkimas	0.154369
	Nėra	0.242283
Triukšmo lygis	Vidutinis	0.363257
	Garsus	0.154824
	Tylus	0.472289
	Labai garsus	-0.165329
Apranga	Įprasta	0.114351
	Prašmatni	0.571225
	Formali	0.139466
Kainos įvertinimas		-0.0454013
Galimybė sėdėti lauke		0.148416
Galimybė rezervuoti staliuką		0.129142
Bevielio ryšio prieinamumas	Nemokamas	0.510277
	Nėra	0.413782
	Mokamas	-0.0990173

Lyginant modelio prognozuojamas reikšmes su realiomis matoma (18 pav.), jog apmokytas tiesinės regresijos modelis geba įvertinti tendencijas: įstaigoms iš testavimo duomenų aibės, kurios turi mažesnį įvertinimas, modelio prognozuojamas įvertinimas taip pat linksta į neigiamą pusę, o reitingui kylant, prognozuojamas reitingas irgi kyla. Tačiau pastebimas ir prognozių išsisklaidymas, ypač įstaigoms su mažesniu reitingu.



18 pav. Tiesinės regresijos modelio prognozių palyginimas su testavimo aibe

Vertinant tiesinės regresijos modelio tikslumą gautos šios paklaidos:

- **Vidutinė absoliutinė paklaida:** 0.595418
- **Šaknis iš vidutinės kvadratinės paklaidos:** 0.729202
- **Reliatyvi absoliutinė paklaida:** 1.017551
- **Reliatyvi kvadratinė paklaida:** 0.920758
- **Užtikrintumo koeficientas:** 0.079242

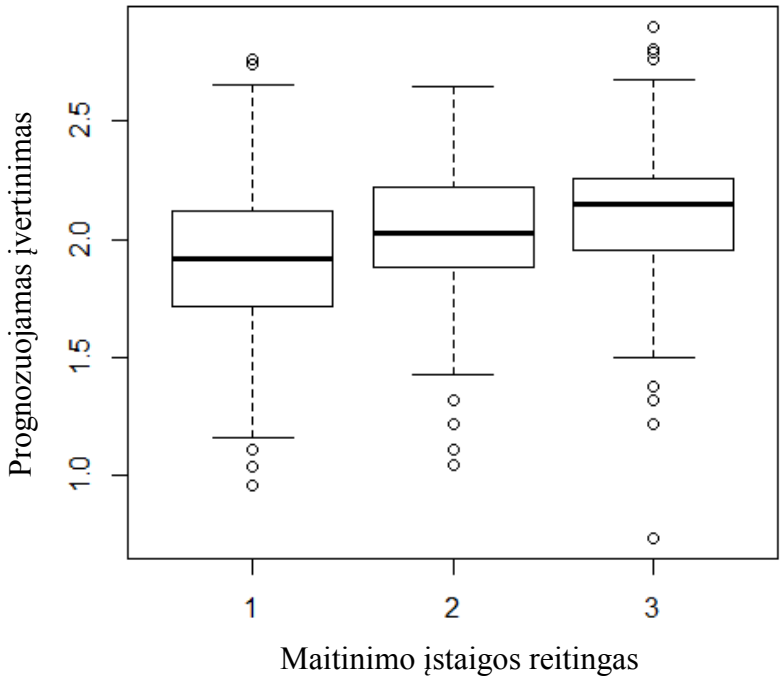
3.3.3. Tyrimas Nr. 3

Šiame tyrime yra apmokomas sprendimų medžio algoritmas. Regresijos medžių sudarymui pasirinktas išplėsto sprendimų medžio metodas (angl. *boosted decision tree*).

Išplėtimas (angl. *boosting*) reiškia, jog kiekvienas medis yra priklausomas nuo prieš tai buvusių medžių ir apsimoko pritaikydamas praeitos iteracijos medį. Todėl šis būdas sprendimų medžių sudarymui didina modelio tikslumą su maža prastesnio modelio padengimo rizika.

Apmokant šį modelį buvo panaudoti sekantys parametrai:

- **Apsimokymo dažnis** – didesnė reikšmė lemia greitesnį konvergavimą, tačiau gali peršokti lokalų minimumą, reikšmė 0.2
- **Maksimalus pavyzdžių kiekis formuojant lapo viršūnę** – reikšmė 10
- **Maksimalus lapų kiekis per medį** – reikšmė 20
- **Sudarytų medžių skaičius** – reikšmė 100



19 pav. Sprendimų medžio modelio prognozių palyginimas su testavimo aibe

Apmokytas sprendimo medžio modelis (19 pav.), kaip ir tiesinės regresijos modelis gebėjo aptikti augimo tendenciją bei vidutiniškai aukštesnį reitingą iš testavimo aibės turinčioms maitinimo įstaigoms priskirdavo aukštesnį reitingą.

Taip pat lyginant pokytį tarp kiekvieno reitingo vidurkio, augimas yra didesnis nei tiesinėje regresijoje. Tačiau atkreipiant dėmesį į minimalias ir maksimalias modelio

prognozuojamas reikšmes, augimo tendencija didėjant tikrajam reitingui pastebima tik minimaliose reikšmėse.

Vertinant sprendimo medžio modelio tikslumą gautos šios paklaidos:

- **Vidutinė absoliutinė paklaida:** 0.619202
- **Šaknis iš vidutinės kvadratinės paklaidos:** 0.750757
- **Reliatyvi absoliutinė paklaida:** 0.997842
- **Reliatyvi kvadratinė paklaida:** 0.946277
- **Užtikrintumo koeficientas:** 0.053723

3.3.4. Tyrimas Nr. 4

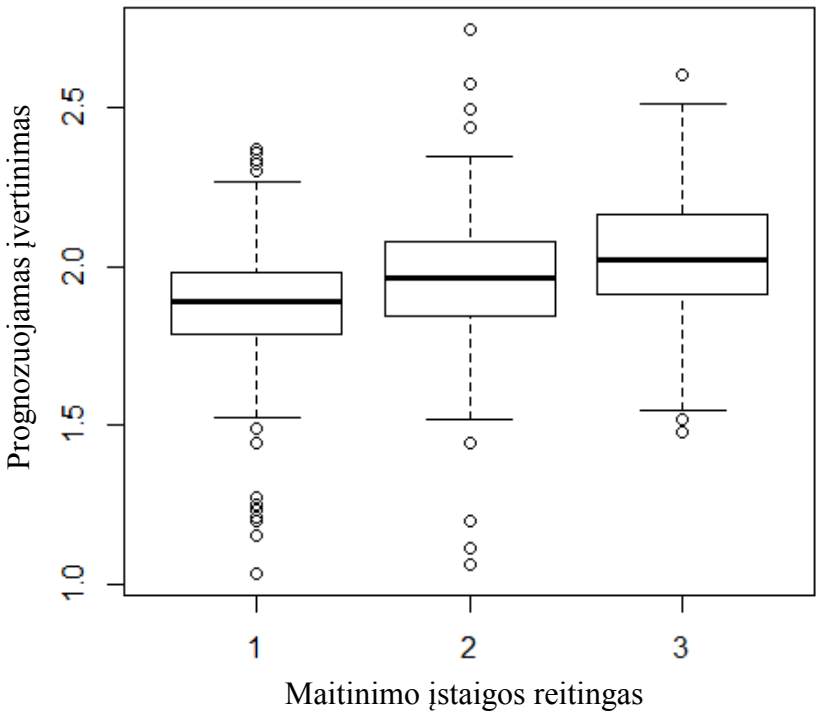
Neuroninių tinklų taikymas plačiai paplitęs sudėtingų problemų sprendimui, kaip vaizdų atpažinimas, nes jie yra lengvai pritaikomi bet kokiai regresiniai problemai. Neuroninių tinklų regresija reikalauja, kad duomenų modelyje būtų tikslinis kintamasis, kadangi regresija grąžina skaitinę reikšmę, tai ir šio kintamojo reikšmė turi būti skaitinė, kaip ir yra maitinimo įstaigų reitingų prognozės modelyje.

Neuroninį tinklą galima interpretuoti, kaip svertinį kryptinį aciklinį grafą. Grafo viršūnės yra išrikiuotos „sluoksniais“ ir sujungtos svertinėmis briaunomis su sekančiu sluoksniu. Pirmasis sluoksnis vadinamas įvesties, o paskutinis regresijos atveju – išvesties, turintis tik vieną viršūnę.

Likę sluoksniai vadinami paslėptais. Tam, kad apskaičiuoti išvesties reikšmę, duotai įvesčiai, reikšmės yra apskaičiuojamos kiekvienai viršūnei tarp paslėptų sluoksnių ir išvesties sluoksnio. Kiekvienai viršūnei reikšmė apskaičiuojama, vertinant svertinę sumą iš buvusio sluoksnių ir taikant jai aktyvavimo funkciją.

Neuroninių tinklų modelį nusako, jo grafo struktūra. Maitinimo įstaigų reitingo prognozei skirtas neuroninis tinklas aprašytas šiais parametrais:

- **„Paslėptųjų“ sluoksnių skaičius** – reikšmė 150
- **„Paslėptųjų sluoksnių viršūnių sujungimas** – reikšmė pilnai sujungtos
- **Normalizatoriaus tipas** – reikšmė „Min-Max“ tipas
- **Pradinių mokymosi svorių diametras** – reikšmė 0.1
- **Mokymosi greitis –** reikšmė 0.005
- **Mokymosi iteracijų kiekis** – reikšmė 200



20 pav. Neuroninių tinklų medžio modelio prognozių palyginimas su testavimo aibe

Analizuojant neuroninių tinklų apmokyto modelio prognozių rezultatus (20 pav.), galima teigti, jog šis modelis kaip ir tiesinė regresija bei sprendimų medis, sugebėjo aptikti tikslinio kintamojo priklausomybę nuo savybių ir vidutiniškai, mažesnį reitingą turinčioms įstaigoms skiria blogesnį įvertinimą.

Matome, jog didėjant maitinimo įstaigos reitingui auga ir prognozės mediana, maksimumas ir kvartiliai.

Vertinant neuroninių tinklų modelio tikslumą gautos šios paklaidos:

- **Vidutinė absoliutinė paklaida:** 0.60427
- **Šaknis iš vidutinės kvadratinės paklaidos:** 0.735362
- **Reliatyvi absoliutinė paklaida:** 0.973779
- **Reliatyvi kvadratinė paklaida:** 0.907866
- **Užtikrintumo koeficientas:** 0.092134

3.4. Tyrimų Nr. 2, 3 ir 4 rezultatų suvestinė

Įvertinus mašininio mokymosi algoritmus maitinimo įstaigų reitingo prognozei galima teigti, jog modeliai aptinka dėsningumą, nors jų tikslumas ir nėra didelis. Tačiau jų spėjimus galima naudoti, kaip sudėtingesnės ekspertinės sistemos dalį arba kaip įžvalgų sistemą skirtą maitinimo įstaigų savininkams.

Vertinant algoritmų rezultatus (5 lentelė), galime pastebėti, jog beveik visais atžvilgiais silpniausiai pasirodė sprendimų medis. Lyginant likusius du metodus, matome, jog neuroniniai tinklai dauguma parametrų lenkia tiesinę regresija, kuri dėl savo savybės minimizuoti kvadratinę paklaidą, pateikė keliais procentais geresnius šaknies iš vidutinės kvadratinės paklaidos ir vidutinės absoliutinės paklaidos rodiklius.

Renkantis algoritmą maitinimo įstaigų reitingo prognozavimo ekspertinei sistemai, didžiausią dėmesį reikia atkreipti į užtikrintumo koeficientą, kurio didžiausią vertę gavo neuroninių tinklų modelis.

5 lentelė Mašininio mokymosi modelių rezultatų palyginimas

	Neuroniniai tinklai	Sprendimų medis	Tiesinė regresija
Vidutinė absoliutinė paklaida	0.60427	0.619202	0.595418
Šaknis iš vidutinės kvadratinės paklaidos	0.735362	0.750757	0.729202
Reliatyvi absoliutinė paklaida	0.973779	0.997842	1.017551
Reliatyvi kvadratinė paklaida	0.907866	0.946277	0.920758
Užtikrintumo koeficientas	0.092134	0.053723	0.079242

4. IŠVADOS

1. Tyrimas Nr. 1 parodė, kad maitinimo įstaigų reitingas turi koreliaciją su įstaigos teikiamomis paslaugomis, pavyzdžiui įstaiga suteikianti nemokamą bevielio ryšio prieigą yra linkusi gauti geresnius įvertinimus, nei teikianti ją kaip mokamą paslaugą ir pan. Todėl yra galimybė sudaryti tiesinės reitingo priklausomybės nuo įstaigos teikiamų paslaugų modelius ir sukurti ekspertinę sistemą.
2. Tyrimas Nr. 2 parodė, kad mašininio mokymosi tiesinės regresijos algoritmas gebėjo aptikti tiesinę reitingo priklausomybę nuo maitinimo įstaigų reitingo duomenų modelio savybės tipo elementų ir rasti jų svorius.
3. Tyrimas Nr. 2 parodė, kad tiesinės regresijos modelis su 0.595418 vidutine absoliutine paklaida ir 0.079242 užtikrintumo koeficientu geba prognozuoti maitinimo įstaigos reitingą, pagal jos teikiamas paslaugas.
4. Tyrimas Nr. 3 parodė, kad sprendimų medžio algoritmo sudarytas netiesinis modelis, gali prognozuoti maitinimo įstaigos reitingo kitimo tendenciją (su 0.619202 vidutine absoliutine paklaida ir 0.053723 užtikrintumo koeficientu, atitinkamai 4% ir 32% blogiau už tiesinę regresiją).
5. Tyrimas Nr. 4 parodė, kad mašininio mokymosi sudarytas neuroninio tinklo modelis leidžia prognozuoti maitinimo įstaigų reitingą, pagal nurodytas įstaigos savybes, su 0.60427 vidutine absoliutine paklaida ir 0.092134 užtikrintumo koeficientu, atitinkamai 2% blogiau ir 17% geriau nei tiesinė regresija
6. Tyrimai Nr. 2, 3 ir 4 parodė, jog mašininio mokymosi regresijos metodai leidžia prognozuoti maitinimo įstaigų reitingus ir jų kitimo tendencijas, netgi iš labai abstrakčios informacijos apie tiriamą įstaigą, neįtraukiant tekstinių lankytojų komentarų ar informacijos apie maisto kokybės. Užtenka žinoti pagrindines įstaigos teikiamas paslaugas, kaip bevielio ryšio prieinamumas, galimybė rezervuoti staliuką ir pan., ir algoritmas gali nusakyti numanomą įstaigos reitingą.
7. Tyrimai Nr. 2, 3 ir 4 parodė, kad apmokytus tiesinės regresijos, sprendimų medžio ar neuroninių tinklų modelius, nepaisant nedidelio užtikrintumo koeficiento (atitinkamai 0.079242, 0.053723 ir 0.092134), galima naudoti maitinimo įstaigų reitingų tendencijų prognozavimo ekspertinės sistemos kūrimui arba kaip dalį išsamesnės (turinčios daugiau parametrų) reitingų prognozavimo sistemos, kadangi tirti modeliai sugeba

nustatyti teigiamus ar neigiamus reitingo pokyčius, priklausomus nuo įstaigos savybių.

8. Tyrimas Nr. 4 parodė, jog neuroninių tinklų modelis prognozuojant maitinimo įstaigų reitingus turėjo aukščiausią užtikrintumo koeficientą (42% didesnį už sprendimų medžio ir 17% didesnį už tiesinės regresijos) ir tiksliausiai atitiko statistinį reitingų modelį, todėl yra tinkamiausias ekspertinės sistemos kūrimui.

5. LITERATŪROS SARAŠAS

S. Owen, Mahout in action, 2011.

1]

C. Bishop, Pattern Recognition and Machine Learning (Information Science and

2] Statistics), 2007.

Y. S. Abu-Mostafa, Learning from Data, 2012.

3]

G. D. F. a. P. H. N. Gupta, „Capturing the stars: predicting ratings for service and

4] product reviews,“ 2010.

T. Hastie, The Elements of Statistical Learning: Data Mining, Inference, and

5] Prediction, 2011.

J.-f. Y. X. Guyon, „On the Underfitting and Overfitting Sets of Models Chosen by

6] Order Selection Criteria,“ *Journal of Multivariate Analysis,* 1999.

G. K. T. a. K. K. Yau, „Predicting electricity energy consumption,“ Department of

7] Management Sciences, City University of Hong Kong, 2005.

P. Harrington, Machine Learning in Action, 2012.

8]

H. C. C.-J. H. W.-H. C. a. S. W. Z. Huanga, „Credit rating analysis with support

9] vector machines and neural networks: a market comparative study,“ College of Management, Chang-Gung University, 2003.

Yelp, „Yelp data set challange,“ [Tinkle]. Available:

10] https://www.yelp.com/dataset_challenge/dataset. [Kreiptasi 10 05 2015].

6. PRIEDAI

6.1. Straipsnis anglų kalba

Predicting restaurant ratings by their attributes

A comparison of linear regression, neural networks and decision tree

Aivaras Čiurlionis
Faculty of Informatics
Kaunas University of Technology
Kaunas, Lithuania
aivarasciurlionis@gmail.com

Justas Šalkevičius
Faculty of Informatics
Kaunas University of Technology
Kaunas, Lithuania
j.salkevicius@gmail.com

Abstract — **Additional restaurants service attributes like providing free Wi-Fi or taking reservations can have positive impact on service quality. However, whether these attributes have any impact on Yelp business star rating? Yelp Academic dataset offers a chance to answer this question using a real data, businesses star rating based on Yelp customers reviews. Furthermore, a prediction of catering star rating becomes possible with a help of machine learning techniques. We used three machine learning approaches for the prediction of restaurants ratings by their attributes: linear regression, neural networks and decision tree. The results showed that various restaurants attributes has negative or positive impacts on user ratings and can be used in statistical modeling. In an empirical application to a Yelp ratings study, the neural networks model showed the best results with linear regression not far behind. A sufficient accuracy of prediction can provide businesses owners with information how to increase quality of catering services.**

Keywords – linear regression, neural networks, decision tree, restaurant ratings, Yelp

Introduction

Solutions for crowd-sourced reviews about local businesses such as Yelp becoming more influential each year. An important part of consumer decision about certain business is 5-stars ranking system also used by Yelp. Research has shown that these ratings have a direct influence on sales [1]. Moreover ratings has even more significant impact on catering business as "an extra half-star rating causes restaurants to sell out 19 percentage points more frequently" [2].

Yelp urges businesses to provide extra information about theirs services as attributes (e.g. attire, price range, Wi-Fi availability and etc.). However, one can wonder has these attributes some correlation with business ratings and are having certain positive services like providing free Wi-Fi can positively impacts business rating. Furthermore possibility to predict ratings by set of provided attributes can help restaurants owners to get an insight how adding or changing qualities of certain service could benefit their business.

This paper focuses on using business data from Yelp academic data set with machine learning algorithms to predict ratings based on business attributes. First, we analyzed and filtered raw data to determinate if it would be possible to fit it with statistical models. Second, we have evaluated results and effectiveness of three different machine learning algorithms used for

regression: linear regression, neural networks and decision tree. Finally, we conclude our findings and proposed future works on this topic.

Data representation

Dataset

The data used for our research is Yelp Academic Dataset, which consists data from Phoenix, Las Vegas, Madison, Waterloo and Edinburgh about 42 153 businesses of every description. The data is delivered in json format, each file is composed of a single object type, one json-object per-line. We have used only information about businesses (individual reviews and user information were not included in our research), generally its attributes and a star rating.

For more accurate and peculiar results we have selected businesses related with catering. However, it is not always easy to tell the particular specialization of businesses subject only from the name of the category given in the dataset. Therefore, only those categories which are especially related with catering were suitable: Restaurants, Bars, Lounges, Bowling, Sport bars, Pubs, Cafes, Dance clubs, Gay bars, Wine bars, Karaoke, Cocktail Bars, Cafeteria and Beer Bar. Due to the possible ambiguity, we have decided to exclude some food-related categories (for example, national restaurants, fast food shops, etc.)

A closer look to the dataset showed that not all attributes are provided by every business available, and to have the biggest possible set of data, we had to choose attributes with the most information (the least number of NULL values). The selected values are shown in the TABLE I.

TABLE I The availability of the attributes in dataset.

Attribute	**Availability (%)**
Price range	93
Outdoor seating	83.6
Alcohol	74.6
Attire	74
Reservations	71.6
Noise level	66.8
Wi-Fi	57

To make prediction model more accurate, we have chosen businesses that have all of these attributes. After the selection of suitable categories and attributes, a total of 8000 (~19 % of starting data) businesses have matched these criteria.

However, the distribution of star ratings (Figure I) showed that most of the scores are average (3 to 4 stars) and the numbers of more extreme ratings (1 or 5 stars) are very low. Thus, a prediction model was very likely to give an average score and almost any extreme scores. This has led to a decision to allocate star ratings into three bigger categories: low (1 – 2.5 stars), average (3 – 4 stars) and high (4.5 – 5 stars).

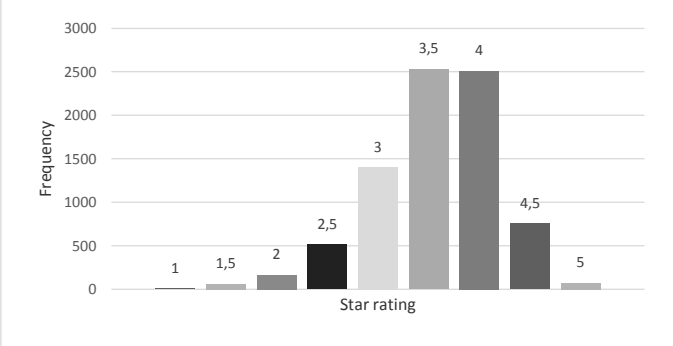


FIGURE I A frequency of star ratings of businesses.

In order to make a prediction model work better, the amount of data in every category had to be similar. To meet these criteria, we have eliminated around 85 % of data from the "Average" category.

TABLE II Frequency of ratings

Rating	**Frequency**	**After elimination**
High	825	825
Average	6434	1265
Low	741	741

Attributes correlation

With Yelp Academic dataset it was possible to determine the attributes impact on overall rating. For example, quiet restaurants mostly have a higher rating than loud ones or that it is better to have a free Wi-Fi than a paid Wi-Fi. Some examples are shown below (Figures II-V).

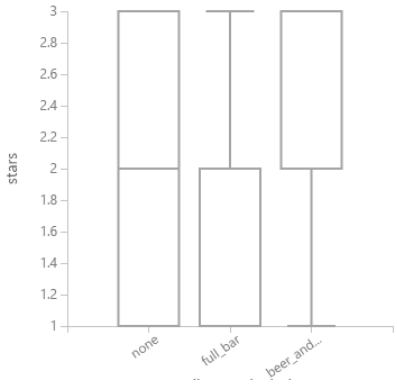


FIGURE II Alcohol

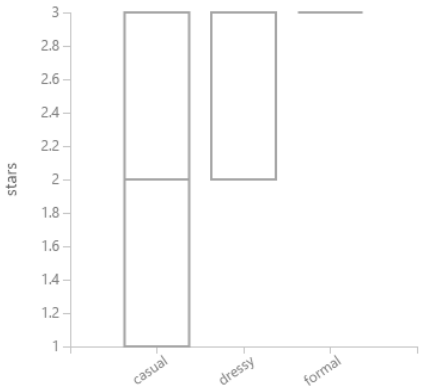


FIGURE IV Attire

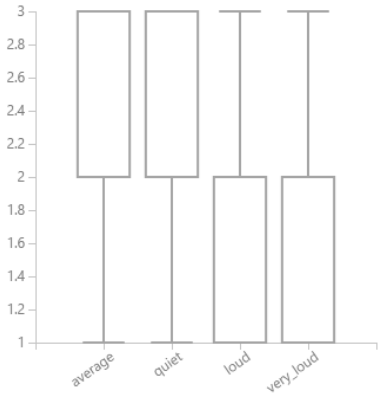


FIGURE III Noise level

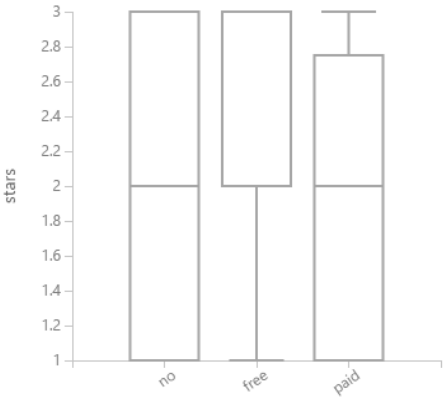


FIGURE V Wi-Fi

Experiments

The most obvious approach to a prediction model is a simple regression. Three different regression models (linear regression, decision tree and neural network regression) were used for predicting a possible rating from business attributes. The accuracy of models were determined by comparing the predictions to actual data form dataset.

Linear regression

Linear regression is an approach for modeling the relationship between a scalar dependent variable y and one or more explanatory variables denoted X. A linear regression model assumes that the relationship between the dependent variable y and vector of regressors xi is linear. In various sets linear regression line might be the same, but the values can be very different. One of its main disadvantages is that if relationship between x and y values is not linear, the results might be very inaccurate. The formula of linear regression:

$$y = a + bx$$

where:

$$a = \frac{\sum y - b\sum x}{n}$$

$$b = \frac{n\sum(xy) - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

x and y are the variables.

b - The slope of the regression line;

a - The intercept point of the regression line and the y axis;

n – Number of values or elements;

In the Table III a predicted attributes impact (positive or negative) to overall rating is shown.

TABLE III Attributes impact on rating

	Feature	**Weight**
Alcohol	Beer and wine	0.42839
	Full bar	0.154369
	None	0.242283
Noise level	Average	0.363257
	Loud	0.154824
	Quiet	0.472289
	Very loud	-0.165329
Attire	Casual	0.114351
	Dressy	0.571225
	Formal	0.139466
	Price range	-0.0454013
	Outdoor seating	0.148416
	Takes reservations	0.129142
Wi-Fi	Free	0.510277
	No	0.413782
	Paid	-0.0990173

According to this model it is way much better to have only beer and wine in your bar than a full bar. The impact of noise level is pretty simple – the quieter your restaurant is, the better. It is important to mention that a high price range gives a negative impact to overall score (when predicting a rating, a value given in the table is multiplied by a price range of business). Another mentionable thing is that a paid Wi-Fi is harmful for restaurant rating.

Although these results look quite logical, this model shows that only a few attributes give a negative impact to the star rating.

This and the following models have been tested with actual data from Yelp Academic dataset. Results are shown in Figure VI.

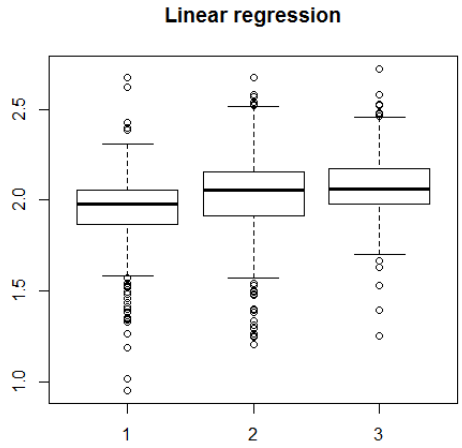


FIGURE VI The linear regression model testing. The model predicted overall rating by business attributes. Horizontal: real rating; Vertical: predicted value.

Decision tree

In decision tree modeling, an empirical tree represents a segmentation of the data that is created by applying a series of simple rules. These models generate set of rules which can be used for prediction through the repetitive process of splitting. After a number of iterations, the final prediction result is chosen.

Results of model testing are shown in Figure VII.

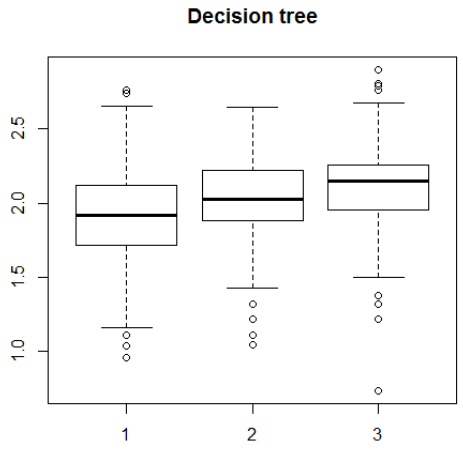


FIGURE VII Decision tree model testing. Horizontal: real rating; Vertical: predicted value.

Neural network regression

Neural network models were originally developed by researchers trying to mimic the neurophysiology of the human brain. Neural networks have been extremely popular for their

unique learning capability and many studies showed that neural networks have performed well in different applications. Artificial neural networks are generally presented as systems of connected "neurons" which can compute values from inputs, and are capable of machine learning.

A results of this model are shown in Figure VIII.

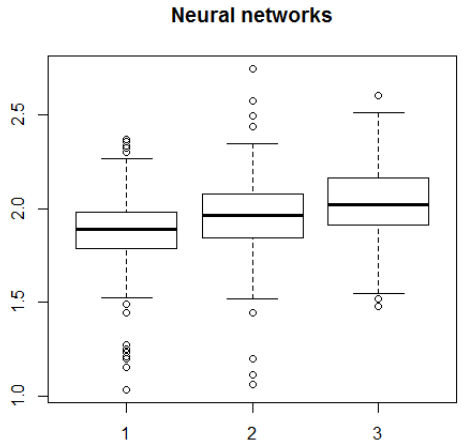


FIGURE VIII Neural networks model testing. Horizontal: real rating; Vertical: predicted value.

Model comparison

All three models predicted a higher value if the actual category is high. On the other hand, the distribution of data was significant. However, as shown in Figure 7, accuracy rates of methods were different (Table IV).

TABLE IV Comparison of models

Model	**Neural networks**	**Decision tree**	**Linear regression**
Mean absolute Error	0.60427	0.619202	0.595418
Root Mean Squared Error	0.735362	0.750757	0.729202
Relative Absolute Error	0.973779	0.997842	1.017551
Relative Squared Error	0.907866	0.946277	0.920758
Coefficient of Determination	0.092134	0.053723	0.079242

The row “Mean absolute Error” shows the average difference between actual and predicted value. In this case, linear regression has the best results, but neural networks are not far behind (2 % worse). The ranking is the same in row “Root Mean Squared Error” - The square root of the average of squared errors of predictions made on the test dataset.

According to “Relative Absolute Error” (the average of absolute errors relative to the absolute difference between actual values and the average of all actual values) row, neural networks have

the smallest error rate and the linear regression has the highest. “Relative Squared Error” (the average of squared errors relative to the squared difference between the actual values and the average of all actual values) also shows the advantage of neural networks model (~ 2 % difference).

The last row shows how well a model fits the data. The closer its value is to one (1.0), the better. Although linear regressions have the lowest mean absolute error rate, neural networks model is ahead in other statistics. Decision tree model results were 42 % worse than neural networks model. And linear regression was 14 % worse than neural networks model.

Results

Our research indicated that some of the attributes has a positive or negative impact on overall business star rating. We found that loud music, paid Wi-Fi and a high price range has a negative impact on star rating, while quiet, formal attire, beer and wine and a free Wi-Fi makes rating higher. However, each business attribute has to be analyzed to determine its significances to rating.

After our research it was clear that various attributes has direct influence on restaurant rating, but it is not the major factor. Three regression models provided similar results, but neural network regression fitted the statistical model the best (42 % and 14 % better than decision tree and linear regression models respectively). All models could predict an adjusted 1-3 star rating with an average absolute 0.6 error rate. So it is significant enough to use it as insights for business owners.

In future, our research could help increase quality of restaurants and other catering services. Owners and managers could get additional information on which attributes of their businesses are the best for the investment in order to maximize attendance of the customers and profit.

Future work

Our model is based only on attributes and star rating, not considering the amount of customer reviews. To make predictions more accurate, we will improve our model with this feature. In addition to this, a bigger database with more attributes is necessary. This will allow predicting, which of them has the biggest impact on star rating.

One of the main problems we encountered was a small number of catering with extreme reviews. More businesses with rating of 1 or 5 stars could provide additional information which attributes are positive for customers and which are negative.

Last but not least, a particular city and its location must be considered. People in different places of the world can be very different and their attitude towards various attributes may vary significantly. In future, we hope to add this feature.

References

M. Luca, *Reviews, Reputation, and Revenue: The Case of Yelp.com,* Harvard Business 1] School, 2011.

M. A. a. J. Magruder, "Learning from the crowd," *The Economic Journal,* 2011. 2]

G. K. Tso and K. K. Yau, Predicting electricity energy consumption, Department of
3] Management Sciences, City University of Hong Kong, 2005.

Z. Huanga, H. Chena, C.-J. Hsua, W.-H. Chen and S. Wu, Credit rating analysis with
4] support vector machines and neural networks: a market comparative study, College of Management, Chang-Gung University, Taiwan, 2003.

N. Gupta, G. D. Fabbrizio and P. Haffner, Capturing the stars: predicting ratings for
5] service and product reviews, Florham Park, NJ 07932 - USA, 2010.

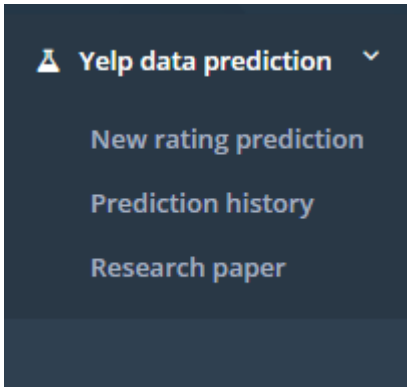
"https://www.yelp.com/dataset_challenge/dataset," Yelp, 2014. [Online].
6]

6.2. Įgyvendintos ekspertinės sistemos vartotojo instrukcijos

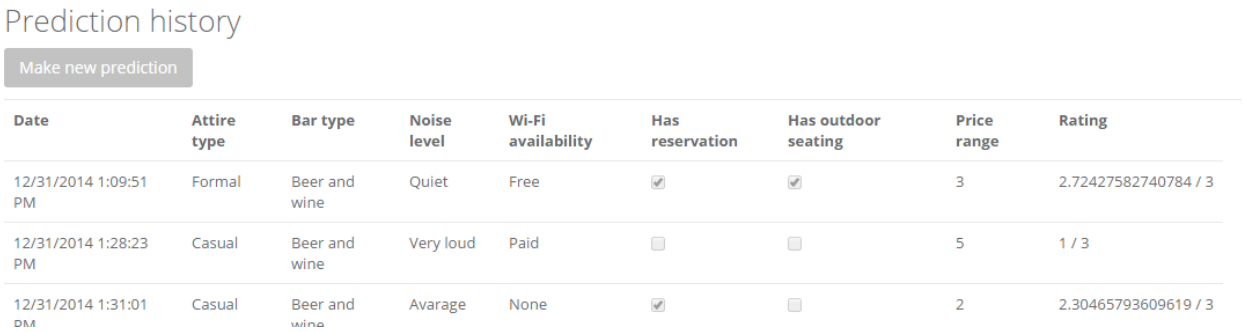
Norėdami pasinaudoti ekspertine sistema apsilankykite:
http://predictionbyattributes.azurewebsites.net/

Kairėje pusėje yra puslapio meniu, kuriame galite:

- Padaryti naują prognozę
- Peržiūrėti prognozių istoriją
- Peržiūrėti mokslinį straipsnį

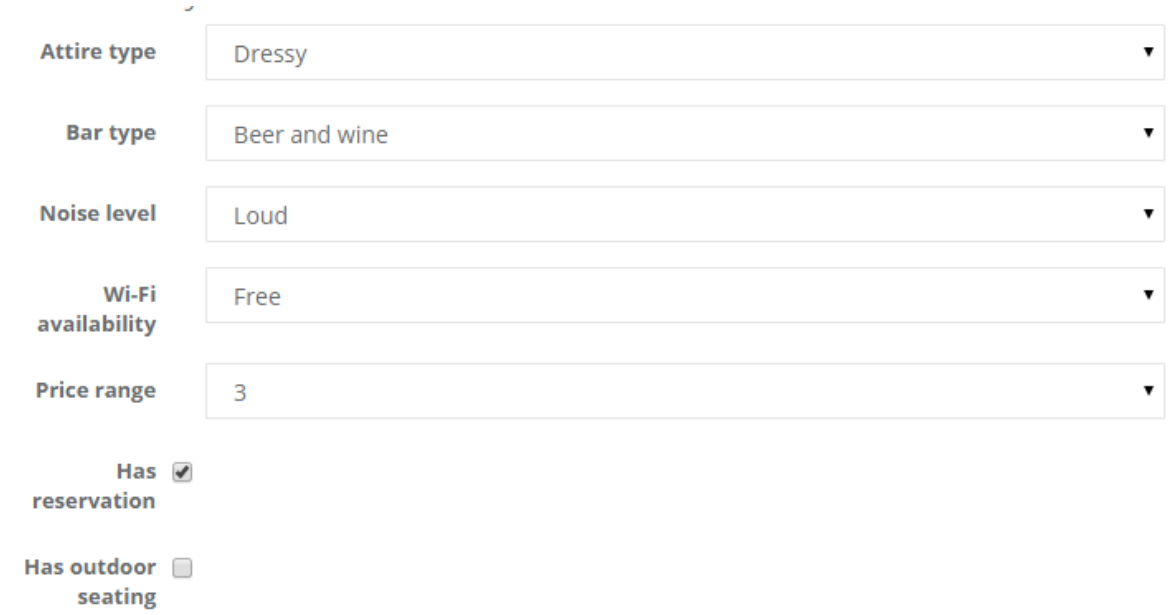


Norėdami padaryti naują prognozę, pasirinkite “Make new prediction”:



Suveskite maitinimo įstaigos, kurios reitingą norite prognozuoti, atributus:

- Pageidaujamos aprangos tipą
- Alkoholio pasirinkimą
- Triukšmo lygį
- Bevielio ryšio prieinamumą
- Galimybę priimti rezervacijas
- Galimybę sėdėti lauke
- Kainos įvertinimą nuo 1 iki 5



Spauskite mygtuką “Create” ir pamatysite prognozuojamą reikšmę: