Predicția prețurilor mașinilor

Gabriel-Ioan Andreica

May 30, 2024

Contents

0.1 Motivația alegerii bazei de date 2
0.2 Contextul bazei de date și al proiectului 3
0.2.1 Cerinte 3
0.3 Obiective 3
0.4 Aspecte teoretice relevante 4
0.5 Implementarea aspectelor teoretice în cadrul proiectului 6
0.6 Testare și validare 8
0.6.1 Metode de testare 8
0.7 Rezultate ale testării 9
0.8 Rezultate 9
0.9 Concluzii 10

Motivația alegerii bazei de date

La fel ca și majoritatea domeniilor, cel al vanzarilor auto, a suferit schimbari majore în ultimii ani, determinate de evoluția tehnologică de la nivel global. Prin digitalizare, mulți dealeri și-au mutat activitatea în online, permițând clienților să vizioneze, compare sau chiar să achiziționeze un autovehicul, direct de pe internet. Această digitalizare, permite, printre altele, dealerilor auto, să colecteze date pe care sa le poată prelucra, cu scopul de a îmbunătăți vânzările. Cu toate că la prima vedere, pare că tehnologia avansată lucrează în folosul firmelor și nu al clientului, aceasta poate fi folosită de ambele părți cu succes. În asemena măsură în care inteligența artificială este folosită în analiza comportamentelor utilizatorilor, aceasta poate fi folosită și în analizarea unui set mare de mașini, pentru prezicerea unui preț corect. Am ales această baza de date, implicit acest subiect pentru proiectul meu, deoarece în perioada alegerii subiectelor, eram în căutarea unei mașini și am vrut să văd dacă mă pot ajuta de cunoștințele mele tehnice în achiziționarea acesteia.

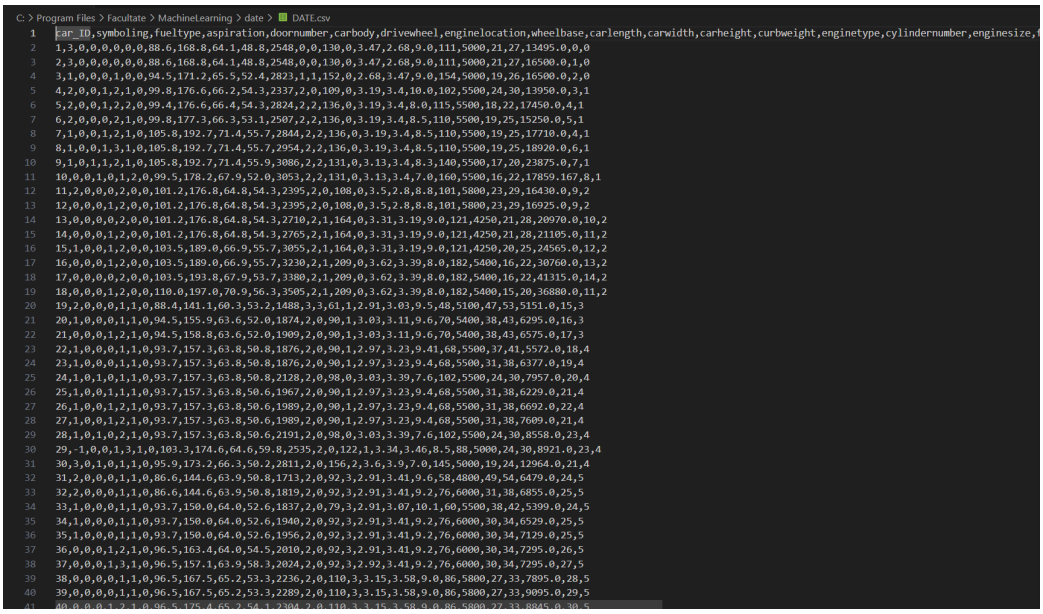

Figure 1: Baza de date pentru test, cu valorile string transformate în numeric.

Contextul bazei de date și al proiectului

Cerinte

Proiectul prezent este un program de Machine Learning care, în baza unui set complex de date, învață să prezică prețul unei mașini. Baza de date cuprinde informații relevante de listă a 205 mașini. Informațiile relevante sunt ID, numele brand-ului, numele modelului, symboling, tipul de combustibil folosit, dacă sunt aspirate sau nu, numărul de uși, tipul de caroserie, tracțiunea, localizarea motorului, wheelbase, lungimea mașinii, lățimea mașinii, înălțimea mașinii, curbweight, tipul de motor, numărul de cilindrii, mărimea motorului, sistemul de combustibil, boreratio, stroke, compressionratio, cai putere, peakrpm, consum în oraș, consum pe autostrada și preț. Cu toate că par multe date, acestea sunt suficiente pentru o predicție de bază a prețului unei mașini. Aplicația nu va putea prezice prețuri pentru mașinile second hand și nu va putea diferenția două mașini identice, dar cu specificații diferite ce țin de confort. Cerințele proiectului sunt să analizeze setul de date, ca mai apoi să extragă din acesta atributele cele mai relevante în contextul prețului unei mașini. În cele din urmă, programul va fi antrenat pentru a putea să fie capabil, ca folosindu-se de acele date care sunt cele mai relevante, să prezică prețul real al unor mașini despre care știe doar specificațiile tehnice, fără preț.

Obiective

Scopul acestui proiect este de a obține în mod automat un preț cât mai apropiat, daca nu chiar exact, de cel real. Aplicația va funcționa doar pentru mașinile noi. Programul va trebui să învețe să interpreteze corelațiile dintre atribute și prețul mașinilor, iar pe baza acestora, va trebui să pună la dispoziție un preț pe care acesta îl consideră ca fiind cel corect pentru vehiculul respectiv. Mai mult decât atât, aplicația va prezice, cu o acuratețe destul de ridicată, prețuri pentru mașini, altele decât cele cu care a fost antrenată. Și pentru acestea, va trebui să estimeze un preț cu o precizie cât mai ridicată, cu toate că este prima dată când programul întâlnește configurația respectivă de atribute. Această aplicație va putea fi folosită atât de persoane fizice, cât și de persoane juridice care vor să achiziționeze o mașină și nu vor să plătească pentru aceasta mai mult decât aceasta merită.

Aspecte teoretice relevante

Inteligența artificială începe să acapareze toate industriile. Modelele de Machine Learning și Inteligență Artificială sunt intens folosite în analiza comportamentală a utilizatorilor. Pe baza acestei analize, companiile de analitică, au tras anumite concluzii și au venit cu soluții sau îmbunătățiri pentru anumite ramuri ale vânzărilor auto. Câteva ramuri pe care eu le consider relevante și în care este folosită inteligența artificială sunt:

- În analiza predictivă [5] a prețurilor. De la bun început, companiile au vânat vânzările rapide, cu prețuri cât mai bune. Pe măsură ce tehnica a avansat, oamenii au început să culeagă informații despre vânzările pe care le-au făcut. Mai târziu, acele informații au stat la baza unor îmbunătățiri care au scăpat unele companii din faliment. În prezent, modelele de Machine Learning pot analiza datele istorice de vânzări, comportamentul clienților și tendințele de pe piață și informațiile pot fi folosite pentru a prezice comportamentul clienților sau pentru a optimiza strategia de vânzări.

- În personalizarea experienței clienților[2]. Orice persoană tinde să cumpere mai mult atunci când se simte bine. Această senzație de bine poate fi dată prin multe moduri, de la o ambianță plăcută în showroom, până la un angajat care îți vorbește frumos și are răbdare cu tine, în alegerea unei mașini pe site. Prin utilizarea procesării limbajului natural, pot fi dezvoltati chatboți care să răspundă la întrebările clienților și să îi ajute în procesul de achiziționare a unui vehicul, precum ar face-o un angajat real, cu experiență în vânzări, psihologie și relații cu publicul.

- În detectarea fraudelor[4]. În paralel cu dorința de a crește vânzările, implicit banii care intră în conturile companiei, s-a dezvoltat și dorința de minimizare a pierderilor. O bună parte din aceste pierderi este dată de persoane rău intenționate care, profitând de anumite vulnerabilități în procesul de achiziție, reușesc să prejudicieze companiile. Machine Learning-ul poate identifica tipare de comportament fraudulos în procesul de finanțare și de achiziționare a unui vehicul prin analizarea datelor [1] tranzacțiilor și a istoricului clienților.

- În previzionarea cererii[3]. Factorul cel mai important în optimizarea vânzărilor, este cererea și oferta. Companiile s-au prins destul de rapid că dacă vor vinde ceea ce nu dorește nimeni să cumpere, nu vor avea

câștiguri prea mari. Evoluția tehnicii a dat o soluție în acest sens. Învățarea automată poate prevedea cererea pentru anumite mașini.

- În campanii de publicitate[2]. Campaniile de publicitate ale companiilor vânzătoare de mașini nu sunt mereu de succes. Unul dintre motive este și lipsa publicului țintă. Inteligența Artificială a venit cu o soluție și în acest sens. Algoritmii de Machine Learning pot analiza comportamentul online al clienților și pot să le afișeze reclame relevante.

Referințele științifice aferente informațiilor de mai sus:

- Dua, S., Du, X. (2019). Data Mining and Machine Learning in Cybersecurity. CRC Press.
- Siegel, E. (2016). Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die. Wiley.
- Chen, T., Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
- Xu, R., Wunsch, D. (2005). Clustering. IEEE Transactions on Neural Networks.
- Phua, C., Lee, V., Smith, K., Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. Artificial Intelligence Review.
- Kouki, Y., Dalle, O., Zouinkhi, A. (2014). Optimization of Inventory Management with Machine Learning. International Journal of Engineering Business Management.
- Dale, R. (2016). The return of the chatbots. Natural Language Engineering.
- Shmueli, G., Patel, N. R., Bruce, P. C. (2010). Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner. Wiley.
- Jardine, A. K., Lin, D., Banjevic, D. (2006). A review on machinery diagnostics and prognostics implementing condition-based maintenance. Mechanical Systems and Signal Processing.
- Fader, P. S., Hardie, B. G. (2009). Probability Models for Customer-Base Analysis. Journal of Interactive Marketing.

Implementarea aspectelor teoretice în cadrul proiectului

Am acordat atentie sporita proiectării și planificării fiecărui task din cadrul proiectului. Acesta a fost împărțit în X pași. Primul pas l-a reprezentat prelucrarea setului de date. După câteva zile în care am analizat datele și structura bazei de date, am ajuns la concluzia că acestea vor trebui prelucrate pentru a fi utilizabile într-un model de Machine Learning. Prima provocare a acestui proiect a fost transformarea valorilor de tip string în valori numerice.

```
def map_non_numeric_values(column):
    unique_values = column.unique()
    mapping = {value: index for index, value in enumerate(unique_values)}
    return column.map(mapping)

# Înlocuirea valorilor non-numerice cu valori numerice pentru fiecare coloană non-numerică
for feature in non_numeric_features:
    df[feature] = map_non_numeric_values(df[feature])
```

Figure 2: Funcția care transformă atributele string în atribute numerice

Pasul următor l-a reprezentat calcularea Gini Index pentru setul meu de date. Gini Index este utilizat în Machine Learning ca o măsură de impuritate pentru a calcula calitatea unui split. În contextul proiectului, este o metrică foarte importantă, pentru alegerea atributelor în nodurile de decizie. După ce am calculat indicele pentru toate campurile din baza mea de date, am ales primele trei cele mai bune caracteristici din punct de vedere al Gini Index. Aceste trei atribute sunt: curbweight, carmodel și carlength.

```
Primii 3 cei mai buni atribuți în funcție de Gini Index:
1. curbweight: 0.0976
2. carmodel: 0.4146
3. carlength: 0.5966
```

Figure 3: Primele 3 atribute din Gini Index

Ulterior, pentru a avea mai multe atribute de calitate pe baza cărora să se efectueze antrenarea modelului, am calculat și Information Quantity. Information Quantity este o măsură a impurității sau dezordinii dintr-un set de date. Această metrică clasifică atributele din setul de date, după relevanța pe care o au în contextul modificării unui atribut vizat. Acest atribut, în proiectul de față este prețul, iar cele trei cele mai bune atribute din punct de vedere al Information Quantity, sunt: enginelocation, aspiration și cylindernumber.

```
Cele mai bune 3 atribute în funcție de cantitatea de informație:
enginelocation: 0.0000
aspiration: -0.0098
cylindernumber: -0.0098
```

Figure 4: Primele 3 atribute din Information Quantity

Pasul următor în proiectul de față, a fost reprezentat de alegerea a 3 modele de Machine Learning pe care ulterior setul de date să fie testat. Cele trei modele alese au fost Random Forest, Decision Tree și Linear Regression. Am efectuat pentru fiecare din cele trei modele o testare cu un set de date care conținea 205 mașini. Pașii precedenți au influențat în mod direct aceste teste. În toate cele trei modele, au fost folosite cele 3 atribute care sunt cele mai bune din punct de vedere al indicelui Gini și cele trei atribute care sunt cele mai bune din punct de vedere al Information Quantity.

Primul model testat a fost Random Forest. Acesta este un algoritm utilizat pentru clasificare și regresie care se bazează pe construirea mai multor arbori decizionali, cu scopul de a obține o precizie cât mai mare. Este avantajos pentru eliminarea problemei overfitting-ului.

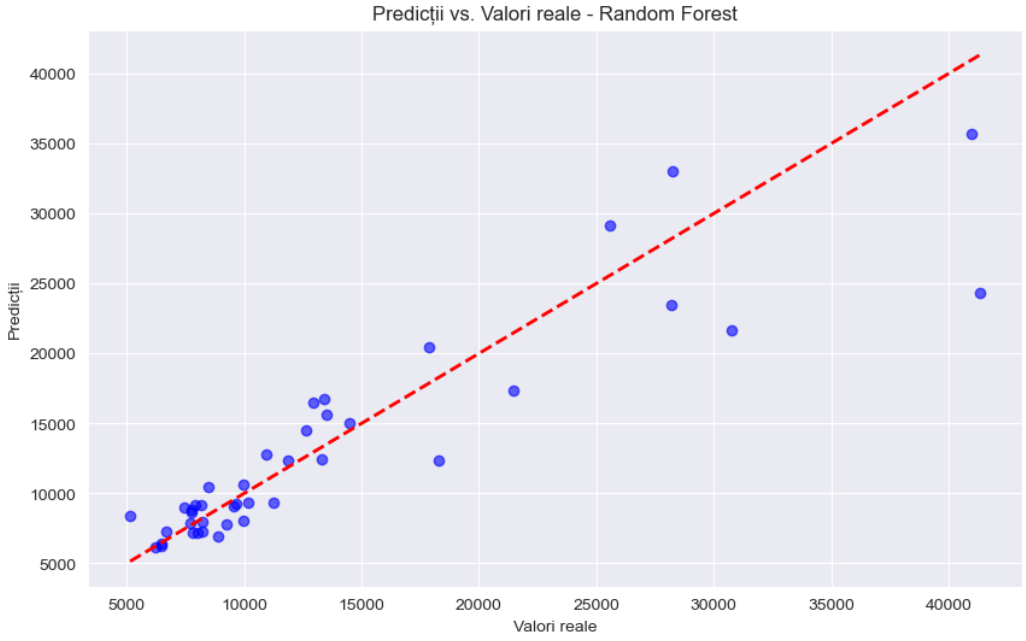

Figure 5: Graficul pentru Random Forest

Cel de al doilea model testat a fost Decision Tree. Asemenea primului model testat, Decision Tree este folosit la clasificare și regresie. Marea diferență între cele două este dată de modul de a funcționa a acestora. De-

cision Tree, spre deosebire de Random Forest, se folosește de o structură de tip arbore în luarea deciziilor.

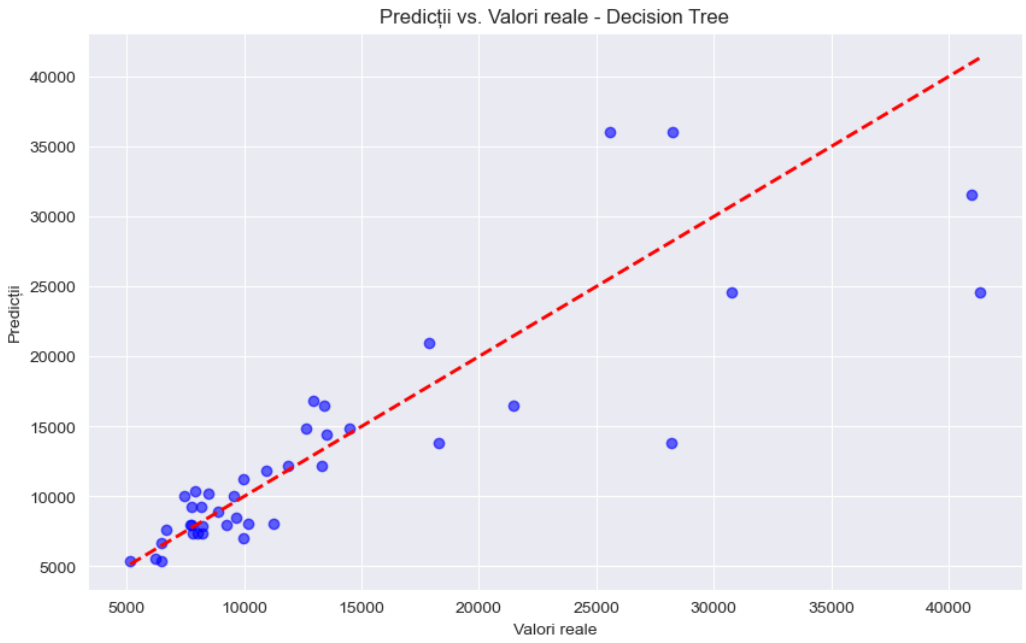

Figure 6: Graficul pentru Decision Tree

Cel din urmă model testat a fost Linear Regression. Acesta este un algoritm utilizat în modelarea relației dintre un atribut cheie și alte câteva atribute.

Dintre toate cele trei modele testate, precizia cea mai bună pe care programul a obținut-o, a fost pe modelul Random Forest, prin urmare, acesta a fost modelul final, folosit în acest proiect.

Testare și validare

Metode de testare

În ceea ce privește testarea modelului acesta de Machine Learning, am urmărit obținerea unei precizii cât mai apropiate de cea pe care am obținut-o pe setul de date de antrenament. Am folosit două variante de testare. În prima, am împărțit setul inițial de date în 70% date de antrenament și 30% date de testare. A doua metoda de testare folosită a fost cea manuală. Am introdus manual într-un nou fișier CSV, specificațiile a 15 mașini, cu diferențe față de

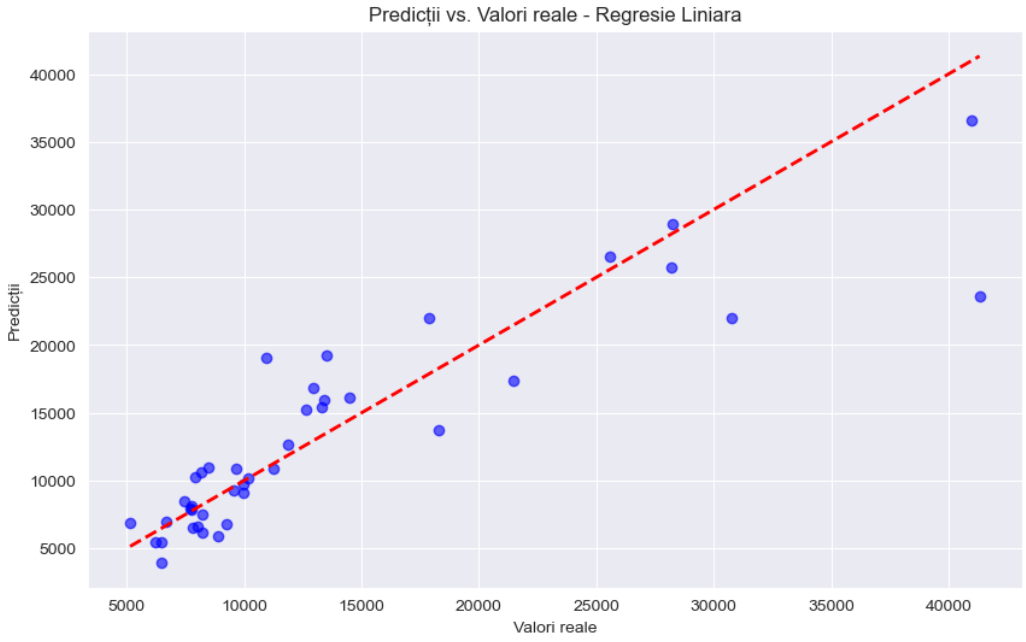

Figure 7: Graficul pentru regresie liniară

cele introduse în setul de date de antrenament. Din nou, atributele au fost transformate din string în numeric, unde a fost cazul și s-a început testarea.

Rezultate ale testării

Rezultatele testării au fost conforme cu așteptările. Atât pentru prima modalitate de testare, cât și pentru a doua, au fost obținute rezultate similare, o precizie în jur de 80-90%. Precizia mai bună a avut-o cea de a doua variantă testată, adică 89.22%. Am observat, de asemenea, că, pe măsură ce prețul vehiculului crește, predicția acestuia este supusă la o eroare mai mare din punct de vedere al diferenței de bani. Acest lucru vine din marja de eroare pe care o mașină de $1000 o poate avea, în contrast cu una de $100000. Cu toate acestea, diferențele sunt din nou asemănătoare, din punct de vedere procentual, raportat la prețul mașinii.

Rezultate

Rezultatele finale ale proiectului sunt conforme cu așteptările pe care le-am impus la începerea acestuia. În ceea ce privește prelucrarea datelor și

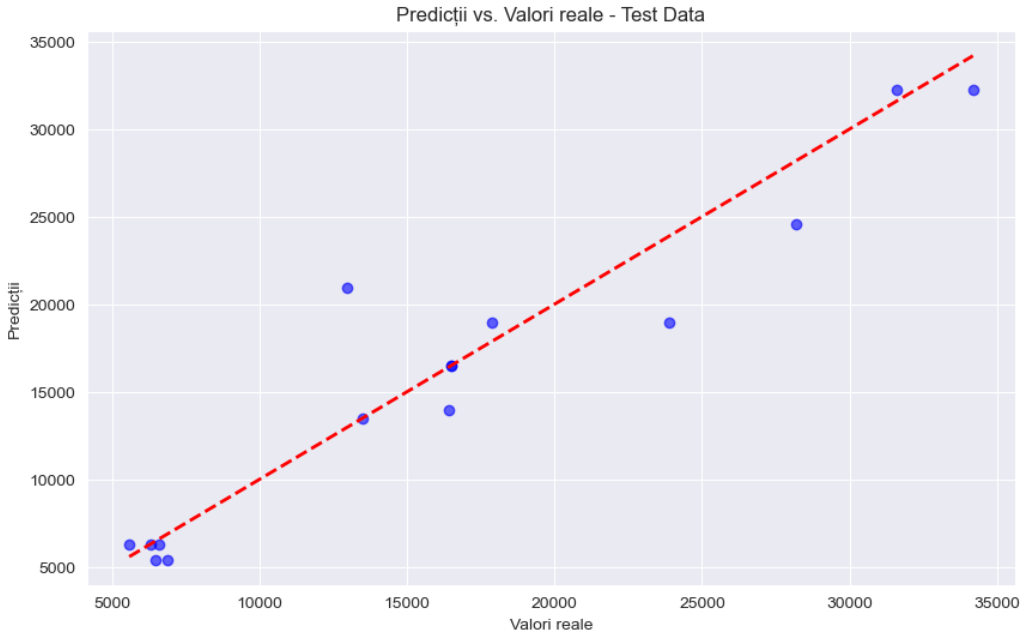

Figure 8: Graficul datelor de testare

antrenarea modelului, metricele rezultate din prelucrarea setului de date au fost de calitate și relevante pentru rezultatul final al proiectului. Mai departe, cele trei modele testate, au fost cele mai bune variante posibile pentru problema dată. Din aceste trei variante, cea mai bună s-a dovedit a fi modelul Random Forest. Chiar și în această situație mai sunt loc de îmbunătățiri. Acestea ar putea fi făcute în partea de prelucrare a setului de date. În loc de transformarea datelor din numeric în string, s-ar putea crea câmpuri noi pentru toate acele date, iar conținutul acestora să fie de tip 0 sau 1, în funcție de valoarea de adevăr a acestora.

Concluzii

Concluziile în urma acestui proiect sunt clare. În ceea ce privește calitățile și pregătirea mea tehnică, am învățat noțiunile introductive din domeniul Machine Learning și Data Analysis. Aceste noțiuni pot fi baza unei cariere în adevăratul sens al cuvântului, pe mai departe, ținând cont de amploarea pe care aceste două domenii o iau. Pe de alta parte, în ceea ce proiectul în sine, rezultatele obținute au fost bune. Ținând cont de acest aspect, putem cataloga proiectul ca fiind de succes. Bineînțeles, programul are loc de mai bine. Acesta ar putea fi îmbunătățit pe viitor. Un exemplu de funcționalitate

pe care programul ar putea-o avea, ar fi aceea de a prezice prețurile și pentru mașinile second-hand, ținând cont de alte atribute, cum ar fi kilometrajul. O alta îmbunătățire ar fi ca programul să țină cont și de atribute ce țin de aspectul sau confortul mașinii, precum opțiunile interioare.