Zadanie 2 - Stromy, stroje, hlasovania a redukcia dimenzie

Pavol Polednák

14. 11. 2023

Contents

1 Predstavenie

V tejto práci sa venujeme vývoju softvérovej aplikácie, ktorej cieľom je predpovedať cenu automobilu na základe rôznych charakteristík. Úloha zahŕňa dôkladné predspracovanie dát, ako je odstránenie nežiaducich a irelevantných informácií, normalizácia dát a aplikácia rôznych metód strojového učenia. Medzi kľúčové aspekty patrí trénovanie a vyhodnocovanie modelov, ako sú rozhodovací strom, ensemble modely a podporné vektorové stroje (SVM), s cieľom dosiahnuť presné a spoľahlivé predpovede. Výsledky modelov sú posudzované na základe metrík ako MSE a R2, pričom dôraz kladieme aj na vizualizáciu a analýzu reziduálov, aby sme zabezpečili hĺbkové pochopenie výsledkov predpovedí.

Ďalšou dôležitou súčasťou nášho projektu je analýza a redukcia dimenzie dát, kde pomocou 3D bodových grafov skúmame vplyv rôznych charakteristík na predpovedanú cenu vozidla. Táto fáza zahŕňa výber vlastností a ich vizualizáciu, s cieľom identifikovať kľúčové vzťahy a závislosti. Následne sa sústreďujeme na výber a optimalizáciu podmnožiny príznakov, kde aplikujeme rôzne metódy ako analýzu korelačnej matice, výber dôležitých príznakov z ensemble modelov a redukciu podľa variancie PCA. Na upravených dátach potom trénujeme modely a porovnávame ich výkonnosť s pôvodnými modelmi.

Riadky datasetu budeme nazývať vzorky(observations/samples) a jeho stĺpce budeme nazývať vlastnosti(features).

2 Predspracovanie údajov

2.1 Definícia údajov

- *ID*: Identifikátor záznamu. Tento stĺpec sa odstraňuje pred odstraňovaním duplikátov.
- *Price (Cena)*: Spojitá číselná hodnota, predstavuje cenu auta v dolároch a je výstupnou hodnotou pre modely.
- *Levy*: Textová hodnota, ktorá sa upravuje na spojitú číselnú hodnotu, označuje odvody.
- *Manufacturer (Výrobca)*: Kategorická hodnota, označuje výrobcu auta.
- *Model*: Kategorická hodnota, označuje model auta.
- *Prod. year (Rok výroby)*: Celočíselná hodnota, označuje rok výroby auta.
- *Category (Kategória)*: Kategorická hodnota, označuje typ vozidla.
- *Leather interior (Kožený interiér)*: Kategorická hodnota, určuje, či má vozidlo kožený interiér.
- *Fuel type (Typ paliva)*: Kategorická hodnota, označuje typ paliva vozidla.
- *Engine volume (Objem motora)*: Hodnota, ktorá sa upravuje na spojitú číselnú hodnotu, označuje objem motora.
- *Mileage (Najazdené kilometre)*: Textová hodnota, ktorá sa upravuje na spojitú číselnú hodnotu.
- *Cylinders (Počet valcov)*: Celočíselná hodnota, označuje počet valcov motora.
- *Gear box type (Typ prevodovky)*: Kategorická hodnota, označuje typ prevodovky.
- *Drive wheels (Pohon auta)*: Kategorická hodnota, označuje typ pohonu auta (predný, zadný, 4x4).
- *Doors (Počet dverí)*: Kategorická hodnota, označuje počet dverí vozidla.
- *Left wheel (Volant naľavo)*: Kategorická hodnota, určuje, či má vozidlo volant naľavo.
- *Color (Farba)*: Kategorická hodnota, označuje farbu auta.
- *Airbags (Počet airbagov)*: Celočíselná hodnota, označuje počet airbagov v aute.
- *Turbo engine (Turbo motor)*: Kategorická hodnota, určuje, či je v aute turbo motor.

2.2 Metodológia úpravy dát

Z datasetu boli odstránené niektoré vlastnosti:

- *ID*: ID je jedinečné pre každý záznam a neobsahuje žiadne informácie, ktoré by boli užitočné pre predpovedanie ceny auta.
- *Model*: Tento príznak sa odstraňuje, pretože obsahuje príliš veľa jedinečných hodnôt, čo vedie k nadmernej veľkosti DataFrame po použití OneHotEncoding. Pri veľkom počte jedinečných hodnôt môže dôjsť k tzv. "prekliatiu dimenzionality", kde model sa stáva menej efektívnym a ťažšie trénovateľným kvôli veľkému počtu vstupných premenných.
- *Color*: Prediktívna sila je nízka v porovnaní s inými, dôležitejšími charakteristikami vozidla.
- *Left wheel*: Tento príznak sa môže odstrániť, ak nie je relevantný pre predpovedanie ceny v konkrétnom trhovom segmente alebo geografickej oblasti.

Niektoré zo spojitých vlastností si bližšie priblížime v nasledujúcej podsekcii. Úprava zvyšných dát prebehla nasledovne:

- Boli odstránené duplikáty.
- Vlastnosti Levy a Mileage boli transformované na číselné hodnoty. Pri Levy boli reťazce "-" nahradené nulou a pri Mileage boli obstránené skratky "km".
- Odstránené vzorky s chýbajúcimi hodnotami pre Levy a Mileage
- Odstránenie hodnôt s Price < 500, a Cylinders < 3 a >12.
- Všetky zvyšné vzorky obsahujúce null boli odstránené. Tento krok je pomerne agrasívny, avšak počet vzoriek v datasete je vysoký a po testovaní sme zistili, že týmto krokov sa už neodstránia žiadne vzorky.
- Pre Price, Levy a Mileage, obsahujúce divoké outliery (viď nasledovná sekcia), sme aplikovali netódu Interquartile Range (IQR) pre ich odstránenie.

Pôvodný rozmer datesetu: (19237, 15)
Nový rozmer datasetu: (14390, 15)

2.3 Univariátna analýza

2.3.1 Price

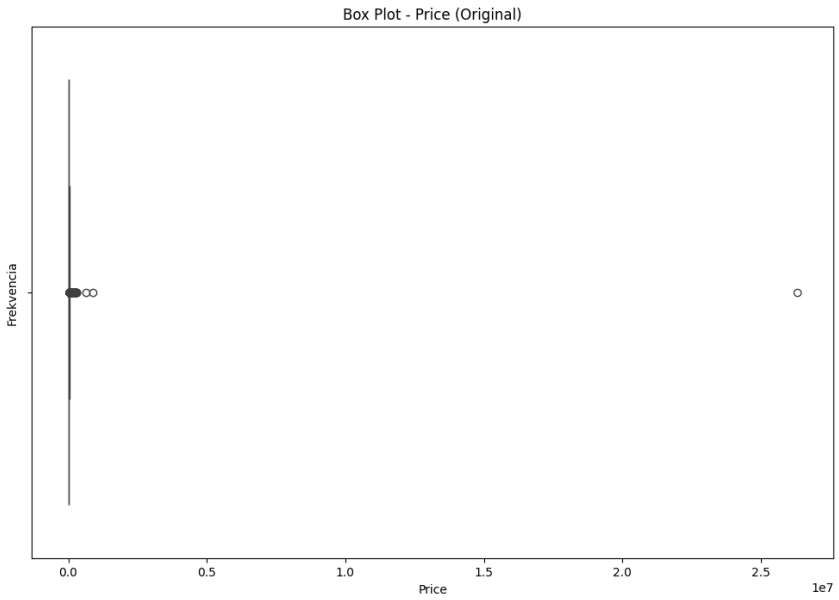

Figure 1: Box plot pre Price - pôvodný

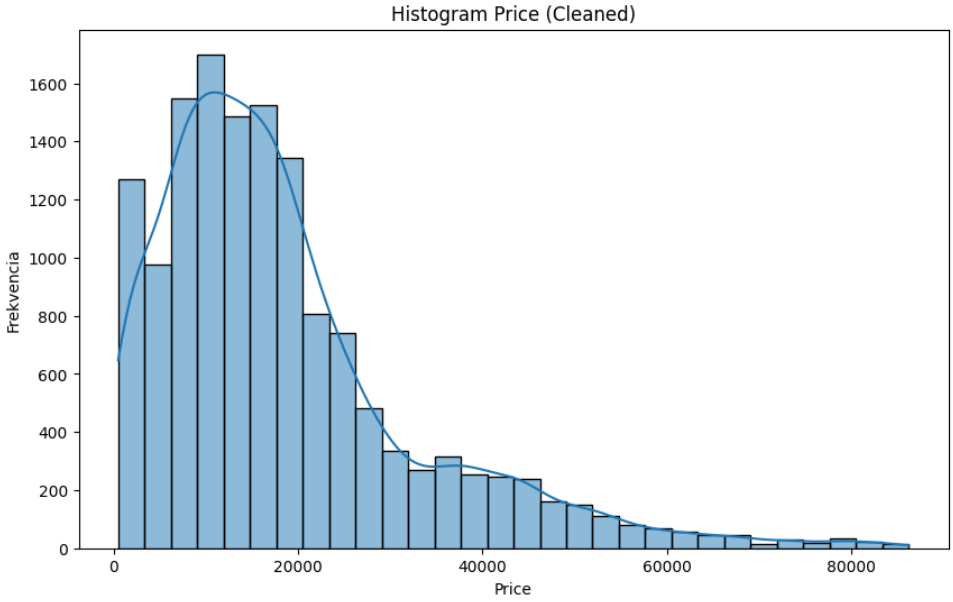

Figure 2: Histogram pre Price - po úprave

Cena obsahuje v pôvodnom datasete vzorky s nereálnymi hodnotami (cena 10 na siedmu - 10 miliónov). Budeme ich považovať za outliery. Po predspracovaní nám ostáva široká škála cien, zahŕňajúce aj ceny hodné špičkových modelov

2.3.2 Levy

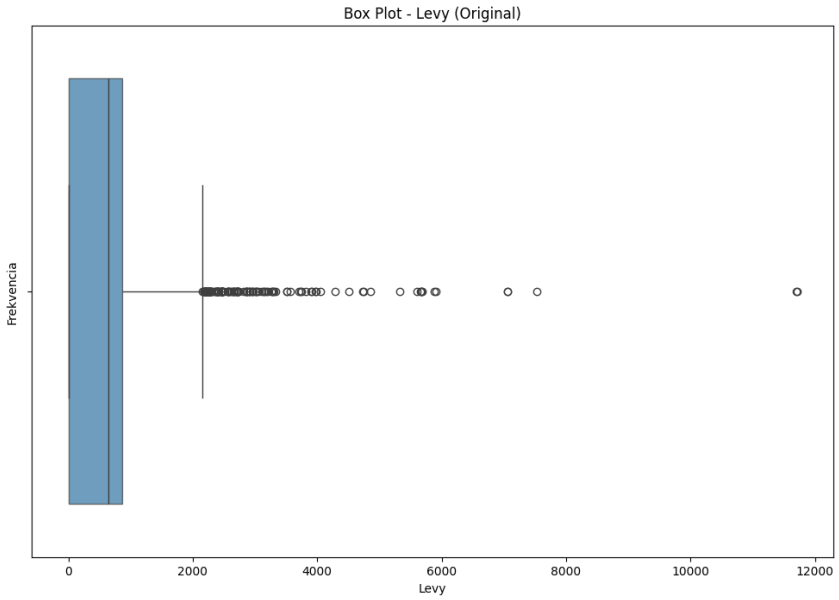

Figure 3: Box plot pre Levy - pôvodný

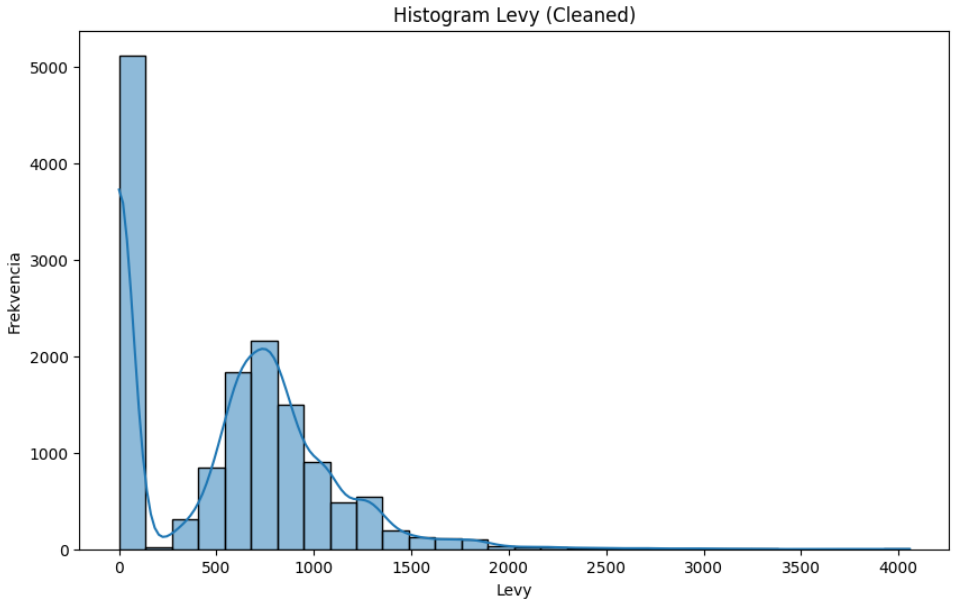

Figure 4: Histogram pre Levy - po úprave

Vidíme, že veľká väčšina hodnôt sa nachádza v rozmedzí 0-2000. existujú však niektoré vzorky s obrovskými hodnotami. Budeme ich považovať za outliery.

2.3.3 Mileage

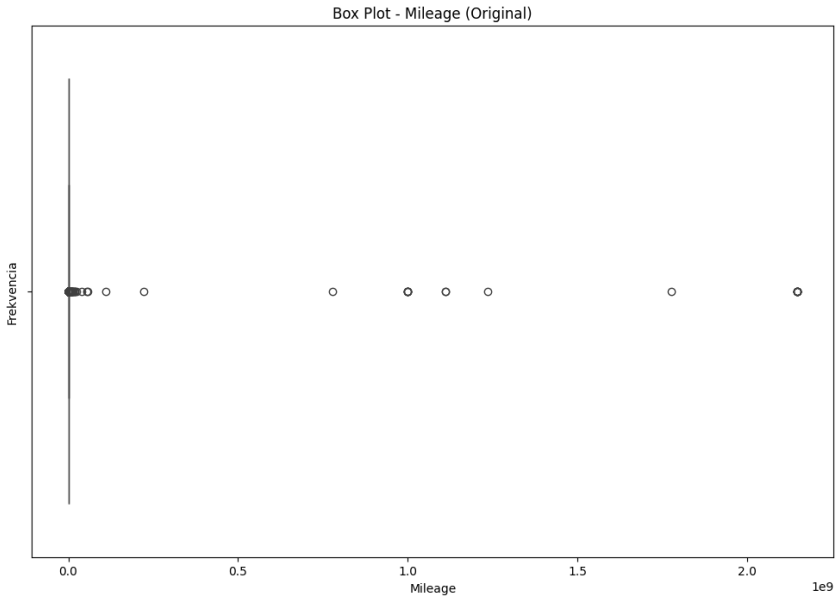

Figure 5: Box plot pre Mileage - pôvodný

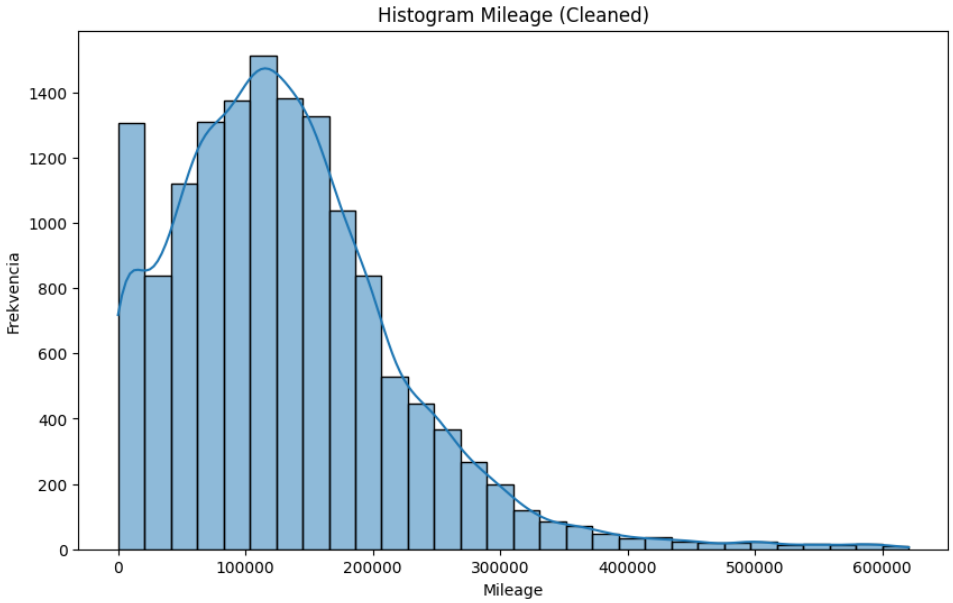

Figure 6: Histogram pre Mileage - po úprave

Pôvodné mileage vzorky obsahujú obrovské hodnoty. Po odstránení výnimiek sa situácia výrazne zlepšila. Vozidlá s mileage nad 500000 zvyknú byť veľmi opotrebované ale rozhodli sme sa tu nebyť príliš restriktívny.

3 Trénovanie modelov

Dataset obsahuje viacero kategorických premenných, ktoré sú pre strojové učenie neinterpretovateľné v ich pôvodnej forme. Keďže kategórie pri našom projekte nemá zmysel zoraďovať, pre ich konverziu do numerického formátu je aplikované one-hot kódovanie. Tento proces vytvára binárne stĺpce pre každú kategóriu v rámci špecifikovaných premenných. Každá unikátna hodnota kategórie sa tak stáva novým prvkom s binárnymi hodnotami, čo zvyšuje kompatibilitu datasetu s algoritmami strojového učenia.

Na vyhodnotenie výkonnosti modelu strojového učenia je dataset rozdelený na trénovaciu a testovaciu množinu. Veľkosť testovacej množiny je nastavená na 10% z datasetu, čo zabezpečuje, že väčšina dát je využitá na trénovanie a zostáva dostatočné množstvo pre vyhodnotenie modelu.

Na záver je vstupná množina normalizovaná pomocou MinMaxScaler z sklearn.preprocessing. Tento nástroj transformuje každý prvok do daného rozsahu, zvyčajne od 0 do 1, čím sa zachováva štruktúra datasetu a zároveň sa zabezpečuje, aby žiadna premenná neprevládala v modeli kvôli jej rozsahu.

Modely budú hodnotené pomocou MSE a R2. Spoločne poskytujú komplexný pohľad na výkon modelu. Zatiaľ čo MSE predstavuje absolútnu hodnotu chyby, R2 predstavuje relatívne hodnotenie, ktoré umožňuje porovnávať modely bez ohľadu na mierku cieľovej premennej. Použitím týchto dvoch metrík môžeme zabezpečiť, že vybrané modely sú nielen presné, ale aj relevantné pre kontext údajov.

3.1 Decision tree

3.1.1 Prvý model

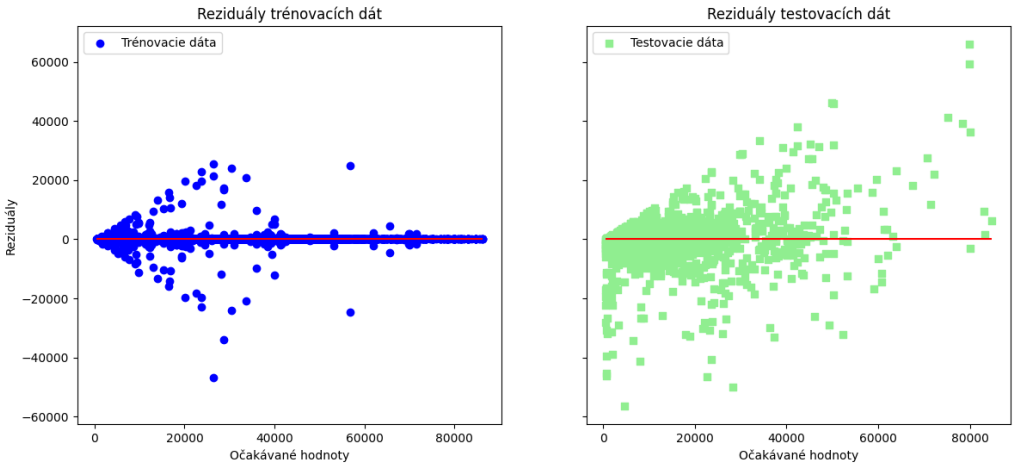

Figure 7: Reziduály pre rozhodovací strom 1

Na úvod sme otestovali základný model rozhodovacieho stromu, bez určenia parametrov:

dt_model = DecisionTreeRegressor(random_state=3)
Výsledkom boli hodnoty pre trénovaciu a testovaciu množinu:
Train MSE: 1223481.3201683268, Train R2: 0.9942208909461568
Test MSE: 84673512.88699713, Test R2: 0.5600197644396312

Trénovacia chyba (MSE) a koeficient determinácie (R2) naznačujú, že model veľmi dobre zapadá do trénovacích dát, s hodnotou R2 (0.9942), čo by mohlo indikovať pretrénovanie na trénovacej množine.
Na testovacej množine dosiahol model R2 skóre 0.56, čo je nad stanovenú hranicu 0.5. Avšak hodnota MSE je pomerne vysoká (84673512.89), čo naznačuje, že predpovede modelu majú značné odchýlky od skutočných hodnôt.
Analýza reziduálov na trénovacích a testovacích dátach 7 odhaľuje niektoré problémy. Reziduály trénovacích dát sú sústredené okolo nuly, čo je indikáciou dobrej prediktívnej schopnosti na trénovacej sade. Naproti tomu, reziduály testovacích dát sú rozptýlené a neukazujú jasný vzor okolo nulovej čiary, čo môže byť znakom pretrénovania modelu, nesprávne špecifikovaného modelu alebo prítomnosti významných odchýlok v dátach.
Odporúča sa ďalšia optimalizácia modelu, napríklad nastavenie maximálnej hĺbky stromu alebo iných parametrov na zabránenie preučeniu. Alternatívne, využitie techník ako sú cross-validácia alebo prístupy k zníženiu dimenzionality môže viesť k zlepšeniu výsledkov na testovacej sade.

3.1.2 Druhý model

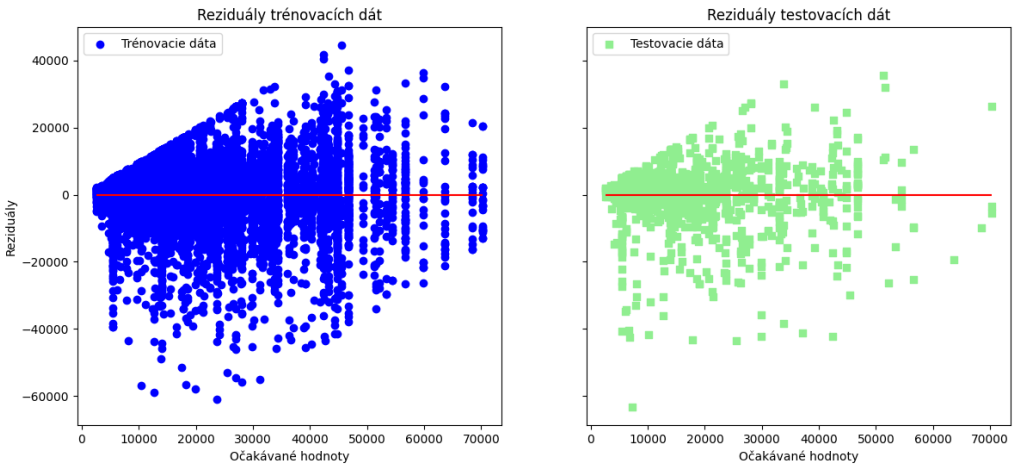

Figure 8: Reziduály pre rozhodovací strom 2

V snahe o zníženie pretrénovania a zlepšenie predikčnej schopnosti modelu bol rozhodovací strom upravený nastavením parametrov. Tieto úpravy boli zamerané na obmedzenie rastu stromu a zvýšenie požiadaviek na rozdelenie uzlov, čo by malo viesť k všeobecnejšiemu modelu:

dt_model = DecisionTreeRegressor(max_depth=10, min_samples_split=40, min_samples_leaf=20, max_features=None, random_state=3)
Výsledkom boli hodnoty pre trénovaciu a testovaciu množinu:
Train MSE: 65781581.46453397, Train R2: 0.6892809667371922
Test MSE: 76515571.16814785, Test R2: 0.6024100349831163

Novo natrénovaný model vykazuje lepšiu rovnováhu medzi schopnosťou predikcie na trénovacej a testovacej sade. Hodnota R2 skóre na trénovacej sade je nižšia (0.6893) v porovnaní s pôvodným modelom, čo naznačuje úspešné zmiernenie pretrénovania. Zlepšenie je viditeľné aj na testovacej sade, kde R2 skóre vzrástlo na 0.6024 a MSE kleslo, čo naznačuje zlepšenú generalizáciu.

Porovnávajúc reziduálne grafy druhého modelu s prvým, je zrejmé, že reziduály sú rozptýlenejšie a menej koncentrované okolo nulovej hodnoty na trénovacej sade, čo je znakom zníženia pretrénovania. Na testovacej sade reziduály vykazujú menšiu variabilitu a sú bližšie k nule, čo ukazuje na zlepšenú prediktívnu schopnosť modelu.

3.1.3 Tretí model

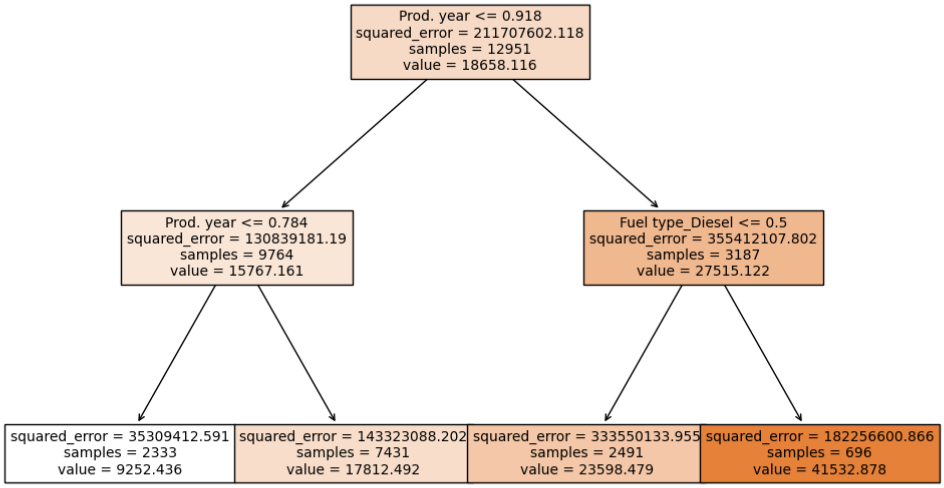

Figure 9: Znázornenie rozhodovacieho stromu

dt_model = DecisionTreeRegressor(max_depth=2, min_samples_split=40, min_samples_leaf=20, max_features=None, random_state=3)
Výsledkom boli hodnoty pre trénovaciu a testovaciu množinu:
Train MSE: 1223481.3201683268, Train R2: 0.9942208909461568
Test MSE: 84673512.88699713, Test R2: 0.5600197644396312

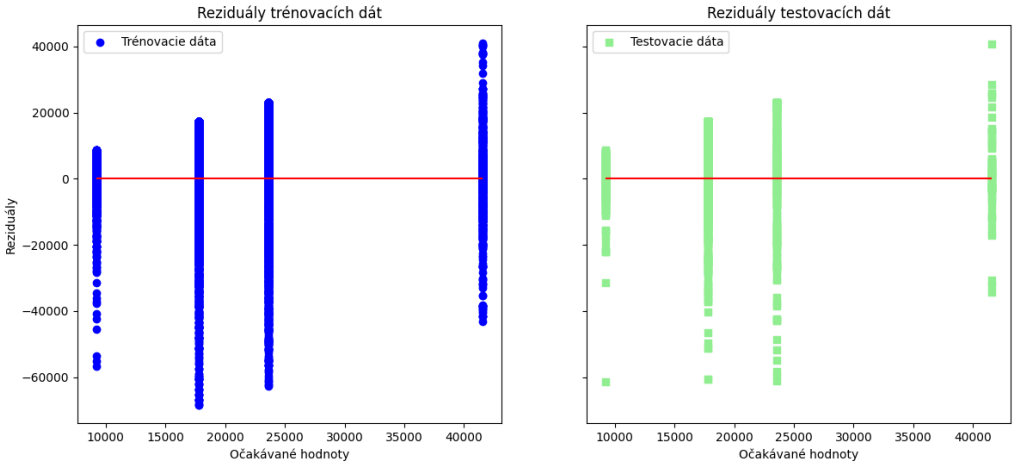

Figure 10: Reziduály pre rozhodovací strom 3

Model rozhodovacieho stromu s obmedzenou hĺbkou bol natrénovaný za účelom vizualizácie a lepšieho porozumenia jeho rozhodovacích pravidiel. Strom s hĺbkou 2 poskytuje jednoduchý prehľad o rozhodujúcich kritériách, ktoré sú základom predikcií modelu. Ako je zrejmé z vizualizácie, hlavné rozdelenie sa uskutočnilo na základe roku výroby ('Prod. year') a typu paliva ('Fuel type'). Táto zjednodušená štruktúra poskytuje rýchly prehľad o dôležitosti týchto prvkov vo vzťahu k cene auta.

Predikčná schopnosť tohto zjednodušeného modelu je podstatne nižšia, čo je očakávané vzhľadom na jeho obmedzenú hĺbku. Hodnoty MSE a R2 skóre sú výrazne horšie ako u predchádzajúcich modelov, čo potvrdzuje, že model s obmedzenou komplexnosťou má nižšiu schopnosť zachytiť variabilitu v dátach.

Reziduálne grafy ukazujú významné rozptyly reziduálov, čo naznačuje, že model nie je schopný presne predpovedať ceny vozidiel a dochádza k veľkým odchýlkam od skutočných hodnôt.

3.2 Random Forest

Model náhodného lesa (Random Forest) bol vybraný ako súborový (ensemble) model, ktorý využíva množstvo rozhodovacích stromov na dosiahnutie lepšej prediktívnej presnosti a stability. Každý strom v lese sa trénuje na náhodne vybranej podmnožine dát s náhodne vybranými vstupnými parametrami, čo znižuje variabilitu a pretrénovanie oproti jednotlivým rozhodovacím stromom.

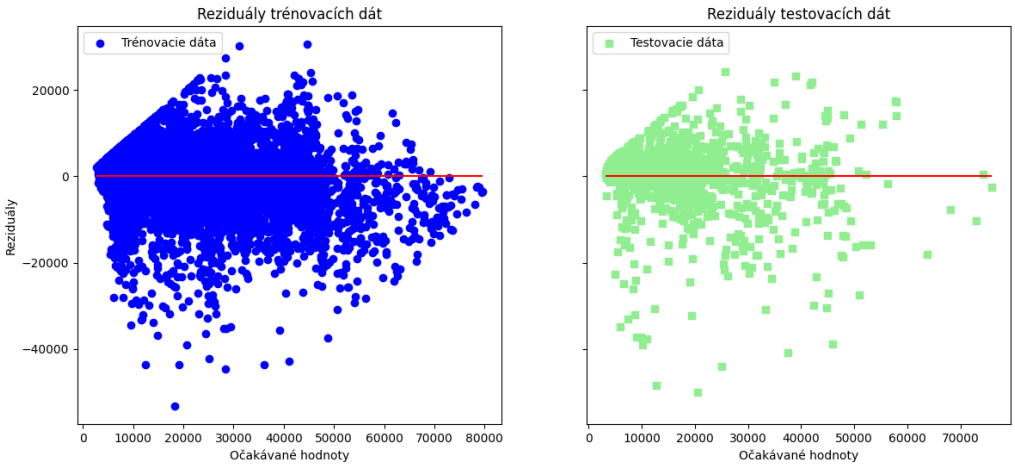

Figure 11: Reziduály pre náhodný les

rf_model = RandomForestRegressor(max_depth=10, random_state=3)
Výsledkom boli hodnoty pre trénovaciu a testovaciu množinu:
Train MSE: 36165881.3778584, Train R2: 0.8291706059866593
TTest MSE: 54811716.30832877, Test R2: 0.7151875358591626

Model náhodného lesa dosiahol na trénovacej sade R2 skóre 0.8292 a na testovacej sade 0.7152, čo predstavuje významné zlepšenie oproti jednoduchým rozhodovacím stromom. Toto zlepšenie je dôsledkom schopnosti náhodného lesa redukovať variabilitu a chyby predpovede prostredníctvom kombinovania predikcií z viacerých stromov. Takto sa znižuje riziko preučenia, a zároveň sa zlepšuje schopnosť modelu generalizovať na nové dáta.

Vizualizácia dôležitosti vstupných parametrov na 12 ukázala, že niektoré prvky, ako sú 'Prod. year', 'Engine volume', a 'Airbags', majú významný vplyv na predpoveď ceny auta. Redukcia prvkov na podmnožinu najdôležitejších môže zjednodušiť model bez výraznej straty informácie a zároveň zlepšiť rýchlosť predikcie a znížiť riziko pretrénovania.

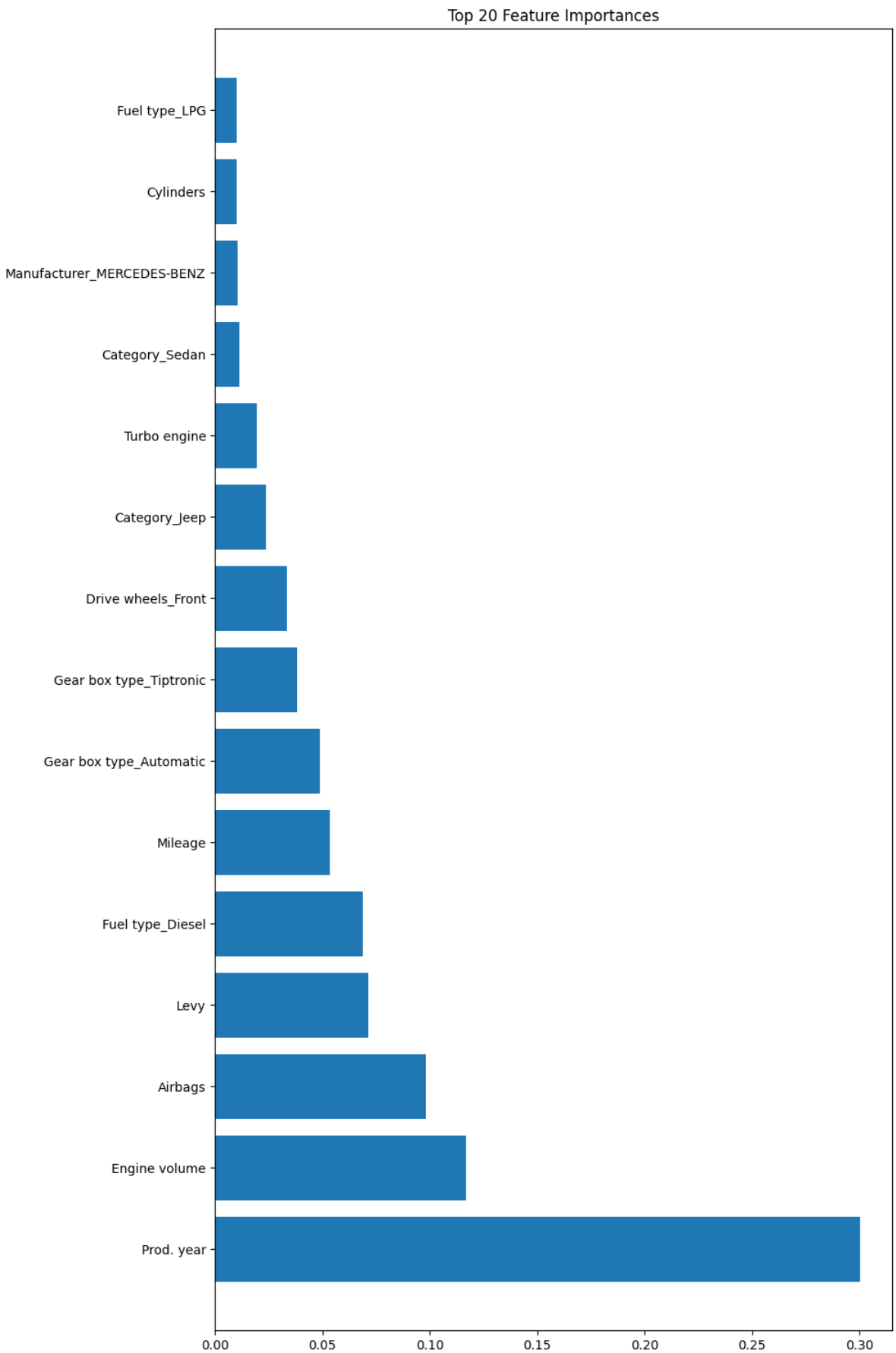

Figure 12: Dôležitosť vstupných parametrov

3.3 Support Vector Machine

Model strojového vektoru podpory (Support Vector Machine - SVM) s polynomiálnym jadrom bol zvolený kvôli jeho schopnosti efektívne pracovať s nelineárnymi vzťahmi v dátach. Iniciálny model SVM s prednastavenými parametrami predviedol negatívne R2 skóre, čo naznačuje horšiu predpovednú schopnosť ako jednoduchý priemerný model.

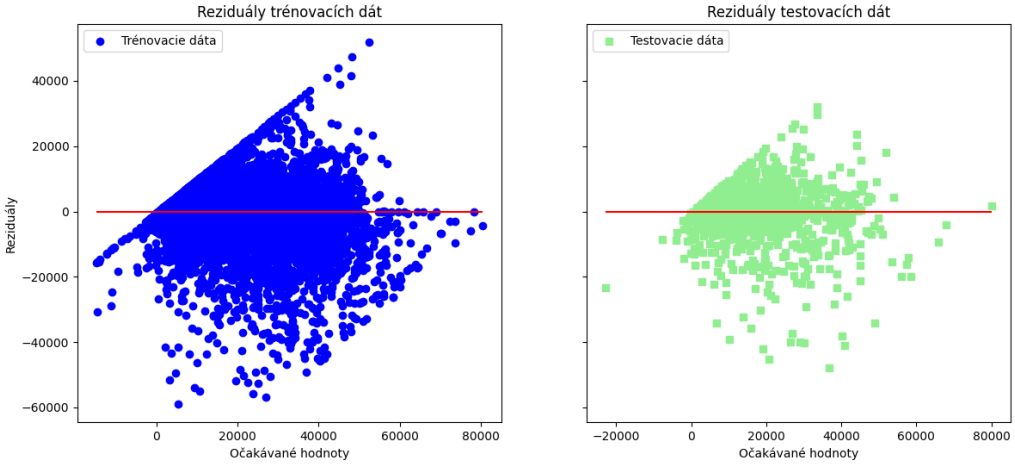

Figure 13: Reziduály pre SVM

Aby sa zlepšila prediktívna schopnosť, bolo vykonané ladenie hyperparametrov. Vybraný bol model s najlepšou dosiahnutou úspešnosťou. Parameter C, ktorý kontroluje mieru regularizácie, bol nastavený na hodnotu 1000, čo znižuje regularizáciu a dovoľuje modelu lepšie sa prispôsobiť dátam. Parameter gamma, ktorý ovplyvňuje tvar rozhodovacej hranice, bol nastavený na 0.5, čo umožňuje modelu identifikovať zložitejšie vzorce. Parameter epsilon bol nastavený na 0.5, určujúc šírku pásmu, v ktorom sa chyby neberú do úvahy v trénovacom procese.

Po ladení hyperparametrov dosiahol model SVM R2 skóre 0.7170 na trénovacej sade a 0.6840 na testovacej sade, čo je značné zlepšenie oproti pôvodnému modelu. MSE hodnoty sú porovnateľné s modelom náhodného lesa, čo indikuje, že model SVM s optimalizovanými hyperparametrami je konkurencieschopný a schopný generalizovať na nových dátach.

4 Redukcia dimenzie a jej vplyv na analýzu

4.1 Ručne vybrané príznaky/vlastnosti

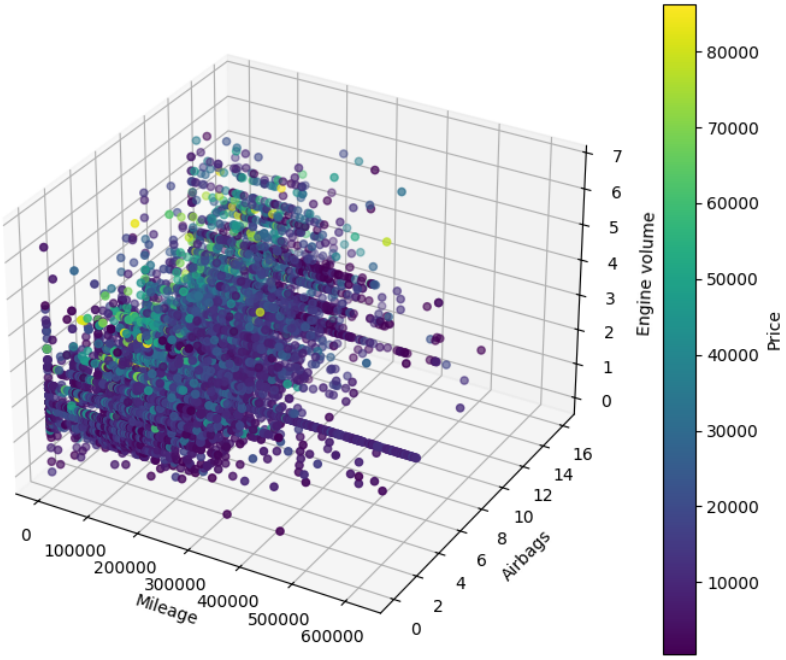

Figure 14: 3D scatter plot pre ručne vybrané vlastnosti

Vybrané vlastnosti: 'Mileage' (najazdené kilometre), 'Airbags' (počet airbagov) a 'Engine volume' (objem motora), kde farebné značenie reprezentuje cenu vozidla. Boli vybrané s predpokladom, že poskytnú prehľad o vzťahu k cene. Z grafu 14 možno pozorovať, že vozidlá s nižším počtom najazdených kilometrov, väčším objemom motora a väčším počtom airbagov majú tendenciu byť drahšie, čo je intuitívne zrozumiteľné a očakávané z hľadiska hodnoty vozidla. Črtajú sa nám taktiež počty airbagov s výrazne viac vzorkami. Hlavnými pravidlami rysujúcimi sa z grafu sú - vozidlá s veĺkým množstvom odjazdených kilometrov budú mať nízku cenu, vozidlá s nízkym objemom motora budú mať nízku cenu.

Druhá časť úlohy využívala techniku (vybrali sme si CPA), aby automaticky redukovala priestor prvkov na tri dimenzie. Na rozdiel od manuálneho výberu PCA nevyberá tri existujúce prvky; namiesto toho vytvára tri nové prvky (hlavné komponenty), ktoré sú lineárnymi alebo nelineárnymi kombináciami pôvodnej sady prvkov, zachytávajúc najviac variancie v dátach. Aj keď konkrétne vzťahy príznakov sú v tomto priestore menej jasné (viď 15), graf môže odhaliť skupiny vozidiel s podobnými cenami a ich vzťahy v zjednodušenom priestore.

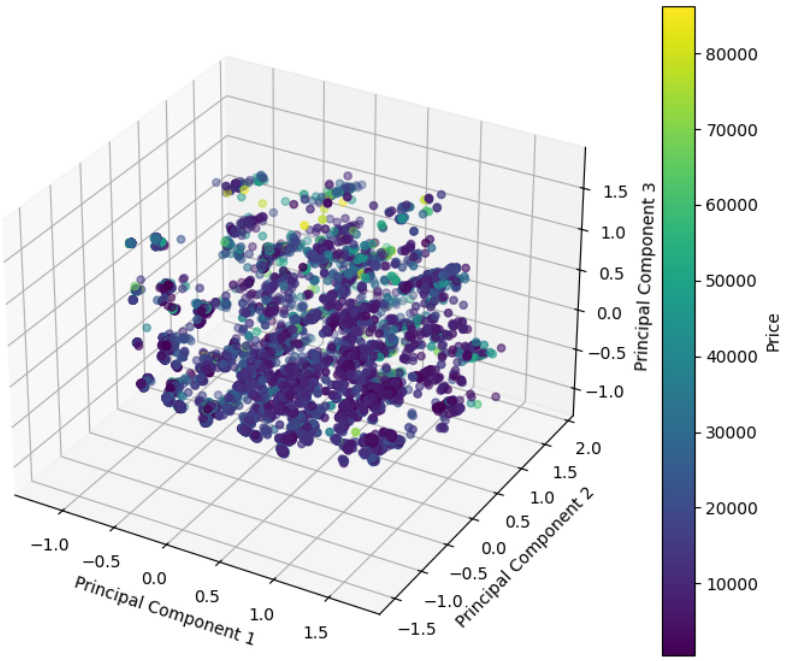

Figure 15: Množina minimalizovaná pomocou PCA

Porovnaním oboch grafov vidíme, že zatiaľ čo prvý graf umožňuje priame interpretácie, druhý graf poskytuje viac abstraktné znázornenie vzťahov. V oboch prípadoch však farebné rozlíšenie podľa ceny umožňuje určité vizuálne oddelenie drahších vozidiel od lacnejších, čo naznačuje, že tieto príznaky sú relevantné pre predpovede ceny.

Porovnanie Metód: Tieto dve metódy sa od seba líšia. Manuálny výber je založený na znalosti domény alebo hypotézach o tom, ktoré prvky sú dôležité a ako by mohli súvisieť s cieľovou premennou. Automatická redukcia spolieha na algoritmické výpočty na nájdenie najlepšej reprezentácie dát v priestore s nižšou dimenzionálnosťou, bez ohľadu na interpretovateľnosť komponentov vo vzťahu k pôvodným prvkám.

Úlohou bolo vykonať obidve aktivity samostatne a potom porovnať výsledné 3D scatter grafy, aby sme videli, ako sa manuálne vybrané prvky porovnávajú s novými dimenziami vytvorenými pomocou PCA. Toto porovnanie pomôže pochopiť, či sa prvky, o ktorých ste si mysleli, že sú dôležité, zhodujú s tými, ktoré algoritmy identifikujú ako zachytávajúce najviac variancie alebo štruktúry v dátach.

5 Trénovanie modelov na zmenšenej množine

V poslednej fáze projektu bola úlohou výber podmnožiny príznakov a opätovné natrénovanie najúspešnejšieho modelu z prvej časti zadania (náhodný les) na zmenšenej množine dát. Tento proces bol vykonaný tromi rôznymi prístupmi:

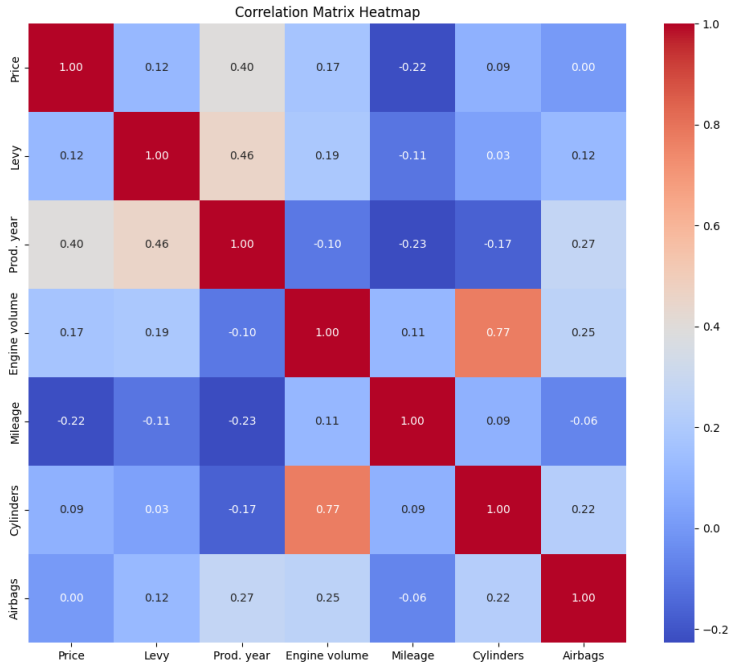

Figure 16: Korelačná matica numbrických dát

Výber Podľa Korelačnej Matice 16: Podmnožina príznakov bola vybraná na základe silnej korelácie s cieľovou premennou 'Price'. Boli vybrané príznaky 'Prod. year', 'Engine volume', a 'Levy'. Náhodný les bol natrénovaný na tejto zredukovanéj sade príznakov. Model dosiahol skóre:

Train MSE: 114647767.71569897, Train R2: 0.4584617341623637
Test MSE: 122277200.86831655, Test R2: 0.3646236017926433

čo naznačuje, že táto zredukovaná sada príznakov nemusí obsahovať dostatok informácií potrebných pre silné predikcie.

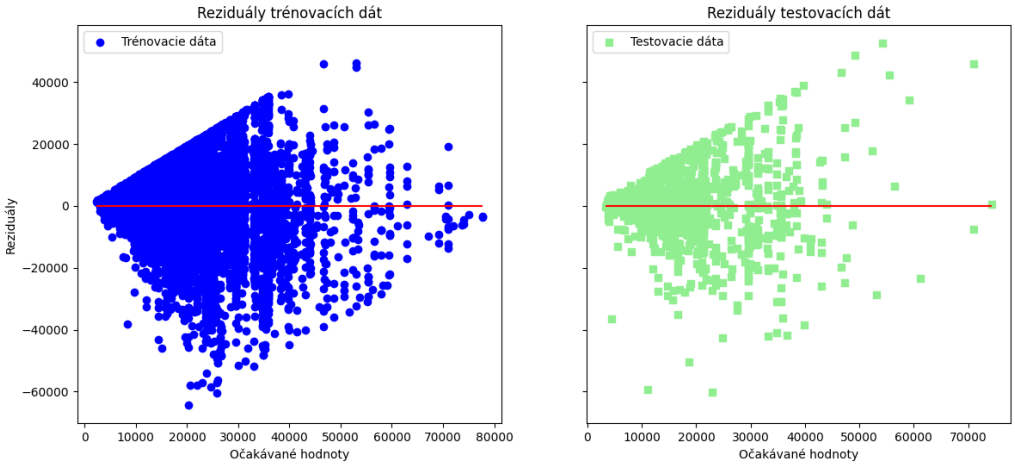

Figure 17: Reziduály na podmnožine príznakov pri výbere podľa korelačnej matice

Výber Podľa dôležitosti príznakov z Ensemble Modelu: Podmnožina príznakov bola vybraná na základe dôležitosti identifikovanej náhodným lesom 12. Táto metóda zahrnula príznaky 'Prod. year', 'Airbags', 'Mileage', 'Fuel type_Diesel', 'Gear box type_Automatic', a 'Category_Jeep'. Po natrénovaní modelu náhodného lesa na týchto príznakoch sme dosiahli skóre:

Train MSE: 67308286.85858436, Train R2: 0.682069580000128
Test MSE: 89785032.58011556, Test R2: 0.5334593022364023

čo je zlepšenie oproti predchádzajúcej metóde. Bolo to spôsobené použitím väčšieho množstva príznakov a ich lepšej schopnosti rozoznávať konečnú vlastnosť.

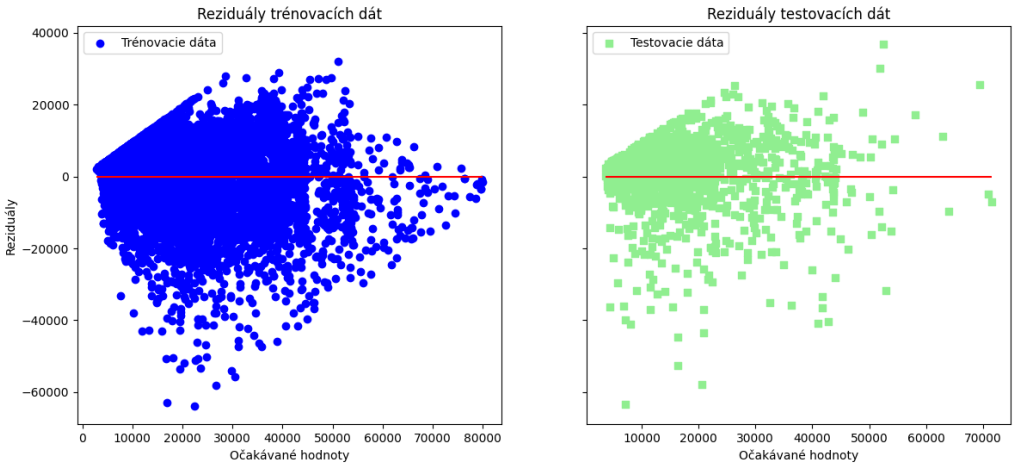

Figure 18: Reziduály na podmnožine príznakov pri výbere podľa dôležitosti

Výber Podľa Variance Pomocou PCA 19: PCA bola použitá na zredukovanie dimenzií vstupných príznakov tak, aby bolo zachované 75% variance dát. PCA transformovala dáta na nový súbor dimenzií, ktorý nezodpovedá žiadnym konkrétnym pôvodným príznakom. Náhodný les bol potom natrénovaný na tejto transformovanej sade. Výsledný model ukázal skóre:

Train MSE: 63033354.948409885, Train R2: 0.7022622035403494
Test MSE: 90321176.50097482, Test R2: 0.5306733929177617

čo poukazuje na relatívnu silu PCA v zachytávaní podstatných vzťahov v dátach.

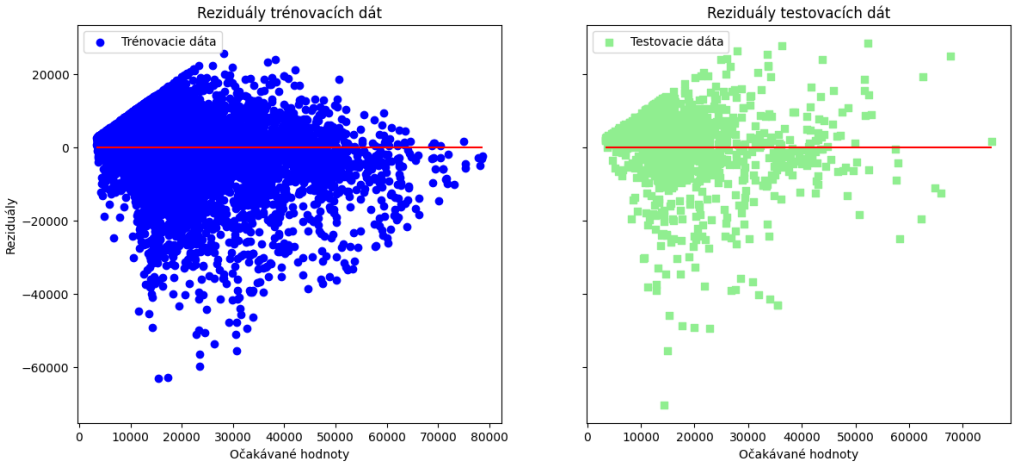

Figure 19: Reziduály na podmnožine príznakov pri použití PCA

Každý prístup má svoje výhody a nevýhody. Zatiaľ čo korelačná matica a dôležitosť príznakov poskytujú priamy pohľad na príznaky a ich vzťah k cene, PCA nám umožňuje zachytiť komplexnejšie štruktúry v dátach, ktoré nemusia byť hneď zrejmé.

Porovnanie Modelov a Reziduálov: Porovnaním výsledkov modelov je jasné, že výber príznakov má významný vplyv na prediktívnu schopnosť modelu. Modely natrénované na podmnožinách príznakov založených na dôležitosti príznakov a PCA dosiahli lepšie výsledky ako model založený len na korelačnej analýze. Reziduálne grafy pre každý model (ktoré tu nie sú podrobne rozobraté) ďalej ilustrujú, ako sa predikcie modelu líšia od skutočných hodnôt a poskytujú nám prehľad o rozptyloch chýb modelu.

V konečnom dôsledku, aj keď výber príznakov pomocou PCA nepriniesol najlepšie skóre R2, poskytol konkurencieschopné výsledky a môže byť preferovaný v situáciách, kde sú interpretovateľnosť a pochopenie vzťahov medzi príznakmi menej dôležité než schopnosť modelu efektívne generalizovať.