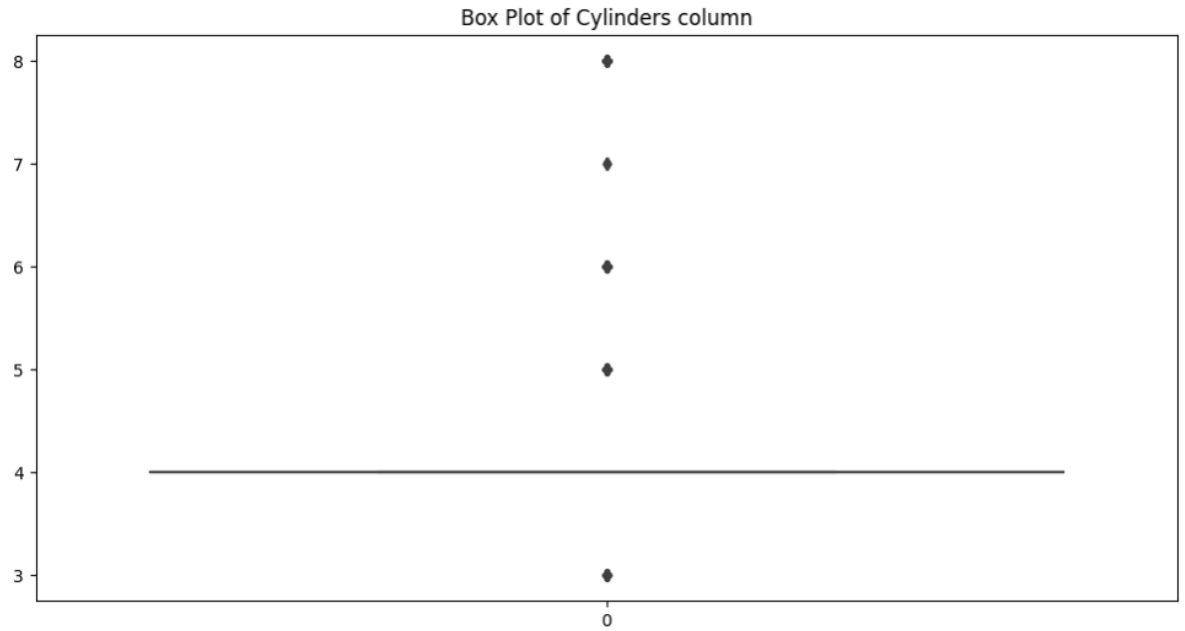
Zadanie 2

Úvod do problematiky

Hlavným zámerom úlohy je vyvinúť funkčný program, ktorý na základe dostupných údajov o aute dokáže predikovať jeho cenu. V tomto dokumente sú popísané kroky, ktoré boli podniknuté pri riešení zadania, a výsledky, ktoré boli dosiahnuté pri riešení jednotlivých podúloh. Súčasťou riešenia je predbežné spracovanie dát po ich dôkladnej analýze, vývoj modelu, trénovanie modelov ako Rozhodovací strom, ensemble model a model SVM (Support Vector Machine), minimalizácia počtu vstupných príznakov, analýza dosiahnutých výsledkov a posúdenie vplyvu zvolených parametrov na tieto výsledky.

○ Odstránenie stĺpcov

Odstránili sme štyri stĺpce. Stĺpce Id a farba som považoval za zbytočné, pretože som nepociťoval, že stĺpec s farbou by pridal významnú hodnotu pri trénovaní. V prípade stĺpca model som nechcel pridať príliš veľa vstupných parametrov, preto som sa rozhodol ho nevyužiť. Čo sa týka stĺpca s počtom valcov, nebol som schopný odstrániť dostatok odľahlých hodnôt bez toho, aby som stratil veľké množstvo dátových záznamov.



V stĺpci levy som zmenil hodnotu '-' na 0 a celý stĺpec som prekonvertoval z typu objekt na int, keďže obsahuje číselné hodnoty.

Stĺpec manufacturer som premapoval na vlastné hodnoty, ktoré reprezentovali krajiny, kde sídlia výrobcovia. Záznamy obsahujúce štyri hodnoty som odstránil, pretože ich počet bol veľmi malý a toto rozhodnutie pomohlo minimalizovať počet vstupných parametrov.

V stĺpci mileage som odstránil reťazec ' km' a zostávajúcu číselnú časť som prekonvertoval z objektu na int.

Hodnoty v stĺpci turbo_engine, ktoré boli číselné, ale mal typ objekt, som tiež prekonvertoval na int.

V stĺpci leather_interior som hodnoty True/False zmenil na číselné hodnoty 1, 0. Z fuel type som odstránil Hydrogen, pretože mal iba jeden záznam, a plugin hybrid som zmenil na Hybrid.

Hodnoty v stĺpci doors som upravil len pre lepšiu čitateľnosť.

Hodnoty v stĺpci left+wheel som prekonvertoval na int.

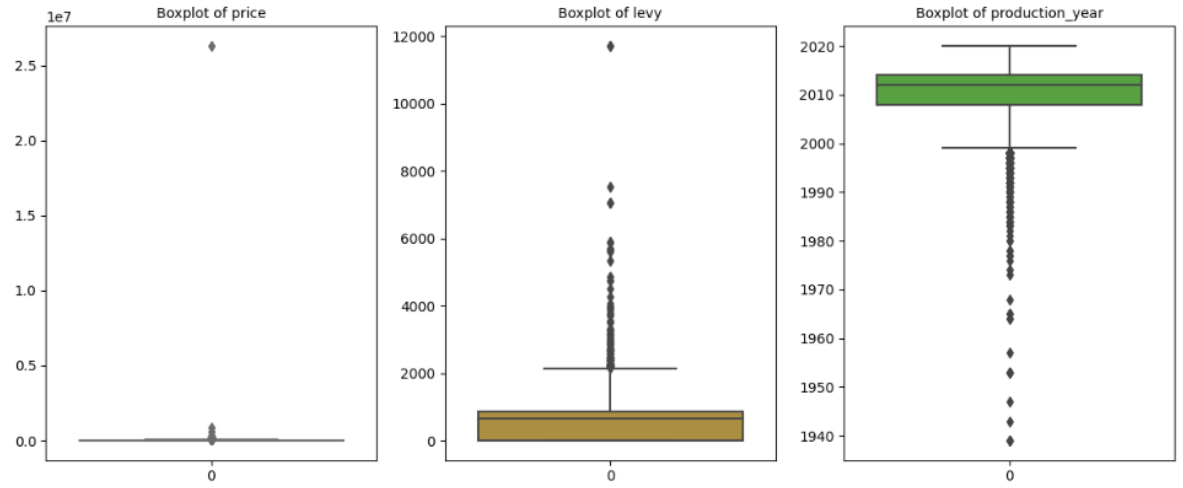
Po odstránení duplicitných hodnôt sa redukoval počet záznamov z 19228 na 15679.

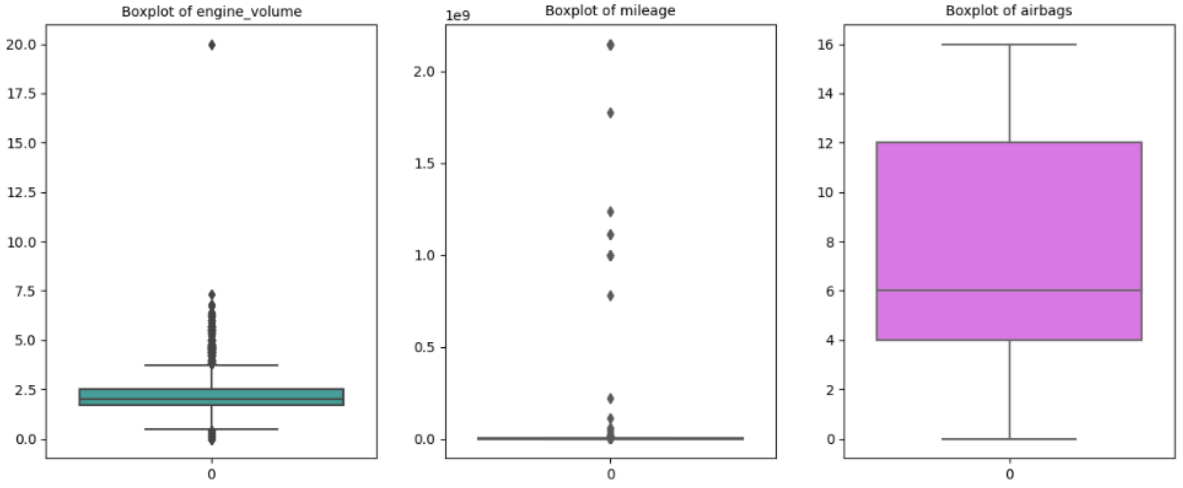
- Odstránenie outlierov

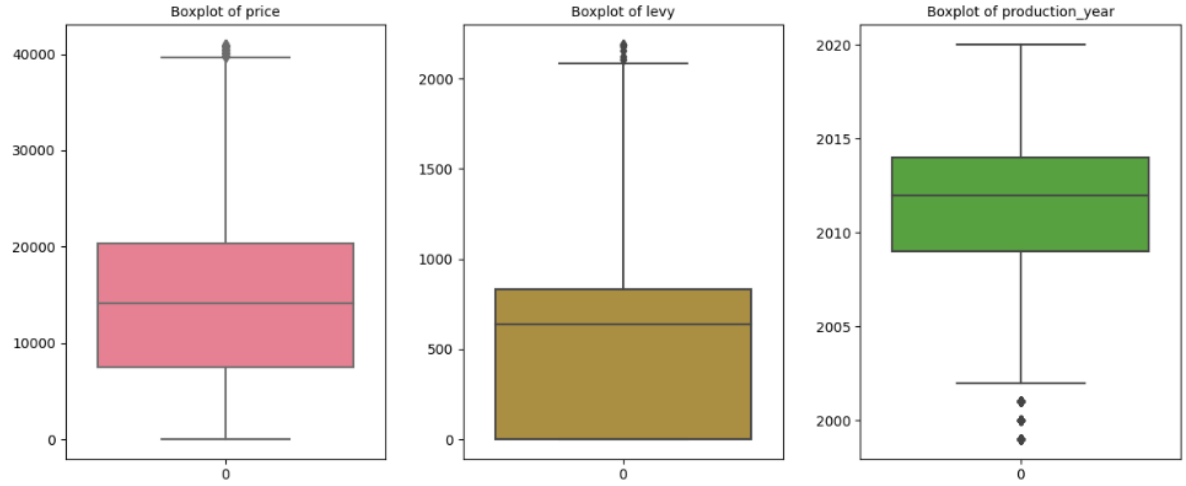
Na nasledujúcom obrázku sú zobrazené boxploty, ktoré ilustrujú odľahlé hodnoty pre jednotlivé stĺpce s číselnými údajmi.

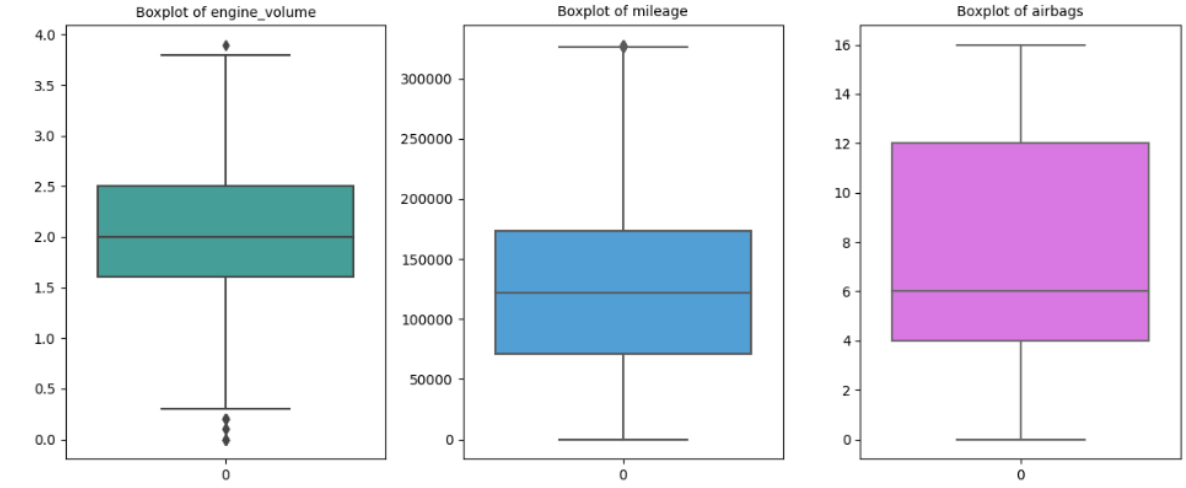
Pokúsil som sa odstrániť odľahlé hodnoty z jednotlivých numerických stĺpcov, aby sme neprerobili viac záznamov v datasete. Na ďaľšej obrázke môžeme vidieť boxploty pre tie isté stĺpce po odstránení odľahlých hodnôt.

Po odstránení vychýlených hodnôt sa redukoval počet záznamov z 15679 na 12462.


Boxplot of price
1e7
2.5
2.0
1.5
1.0
0.5
0.0
0
Boxplot of levy
12000
10000
8000
6000
4000
2000
0
0
Boxplot of production_year
2020
2010
2000
1990
1980
1970
1960
1950
1940
0


Boxplot of engine_volume
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
0
Boxplot of mileage
1e9
2.0
1.5
1.0
0.5
0.0
0
Boxplot of airbags
16
14
12
10
8
6
4
2
0
0


Boxplot of price
40000
30000
20000
10000
0
0
Boxplot of levy
2000
1500
1000
500
0
0
Boxplot of production_year
2020
2015
2010
2005
2000
0


Boxplot of engine_volume
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
Boxplot of mileage
300000
250000
200000
150000
100000
50000
0
0
Boxplot of airbags
16
14
12
10
8
6
4
2
0
0

- One-hot encoding

Pri použití metódy onehot encoding som zakódoval kategorické stĺpce. Tým sa zvýšil počet vstupných parametrov, ktoré môžeme vidieť na nasledujúcom obrázku. Nakoniec budeme pracovať s 45 vstupnými hodnotami.

```
Data columns (total 46 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   price                           12462 non-null  float64
 1   levy                            12462 non-null  int64
 2   production_year                 12462 non-null  int64
 3   leather_interior                12462 non-null  int64
 4   engine_volume                   12462 non-null  float64
 5   mileage                         12462 non-null  int64
 6   airbags                         12462 non-null  int64
 7   turbo_engine                    12462 non-null  int64
 8   left_wheel                      12462 non-null  int64
 9   manufacturer_Czech Republic     12462 non-null  int64
 10  manufacturer_France             12462 non-null  int64
 11  manufacturer_Germany            12462 non-null  int64
 12  manufacturer_Italy              12462 non-null  int64
 13  manufacturer_Japan              12462 non-null  int64
 14  manufacturer_Russia             12462 non-null  int64
 15  manufacturer_South Korea        12462 non-null  int64
 16  manufacturer_Sweden             12462 non-null  int64
 17  manufacturer_UK                 12462 non-null  int64
 18  manufacturer_USA                12462 non-null  int64
 19  fuel_type_CNG                   12462 non-null  int64
 20  fuel_type_Diesel                12462 non-null  int64
 21  fuel_type_Hybrid                12462 non-null  int64
 22  fuel_type_LPG                   12462 non-null  int64
 23  fuel_type_Petrol                12462 non-null  int64
 24  fuel_type_Plug-in Hybrid        12462 non-null  int64
 25  gear_box_type_Automatic         12462 non-null  int64
 26  gear_box_type_Manual            12462 non-null  int64
 27  gear_box_type_Tiptronic         12462 non-null  int64
 28  gear_box_type_Variator          12462 non-null  int64
 29  doors_four_to_five              12462 non-null  int64
 30  doors_more_than_five            12462 non-null  int64
 31  doors_two_to_three              12462 non-null  int64
 32  drive_wheels_4x4                12462 non-null  int64
 33  drive_wheels_Front              12462 non-null  int64
 34  drive_wheels_Rear               12462 non-null  int64
 35  category_Cabriolet              12462 non-null  int64
 36  category_Coupe                  12462 non-null  int64
 37  category_Goods wagon            12462 non-null  int64
 38  category_Hatchback              12462 non-null  int64
 39  category_Jeep                   12462 non-null  int64
 40  category_Limousine              12462 non-null  int64
 41  category_Microbus               12462 non-null  int64
 42  category_Minivan                12462 non-null  int64
 43  category_Pickup                 12462 non-null  int64
 44  category_Sedan                  12462 non-null  int64
 45  category_Universal              12462 non-null  int64
dtypes: float64(2), int64(44)
```

- Rozdelenie dát na trenovaciu a testovaciu množinu

Predspracovaný a vyčistený dataset som rozdelil na trénovací a testovací dataset v pomere 8:2. Trénovaciu časť som vhodne zakódoval pomocou metódy standard scaling. Táto metóda je dôležitá, pretože zabezpečuje, že všetky vstupné premenné majú rovnaké škálovanie, čo je nevyhnutné pre mnohé algoritmy strojového učenia, aby správne fungovali. Bez rovnakého škálovania by mohli mať niektoré premenné neúmerne veľký vplyv na výsledok modelu, čo by mohlo viesť k nesprávnym predikciám.

- Trénovanie rozhodovacieho stromu

Natrénoval som rozhodovací strom s šiestimi rôznymi hyperparametrami. Najlepší model mal nasledujúce hyperparametre: 'max_depth': 10, 'min_samples_leaf': 4, 'min_samples_split': 40. Na trénovacom datasete som dosiahol R2 skóre 0.65, s chybou MSE (Mean Squared Error) 34,123,009.31 a RMSE (Root Mean Squared Error) 5,841.49. Na testovacom datasete som dosiahol R2 skóre 0.61, MSE 37,321,020.64 a RMSE 6,109.09.

Dataset	R2 Score	MSE	RMSE
Training	0.65	34,123,009.31	5,841.49
Testing	0.61	37,321,020.64	6,109.09

Po natrénovaní modelu sme vyhodnotili aj reziduály pre trénovaciu aj testovaciu množinu.

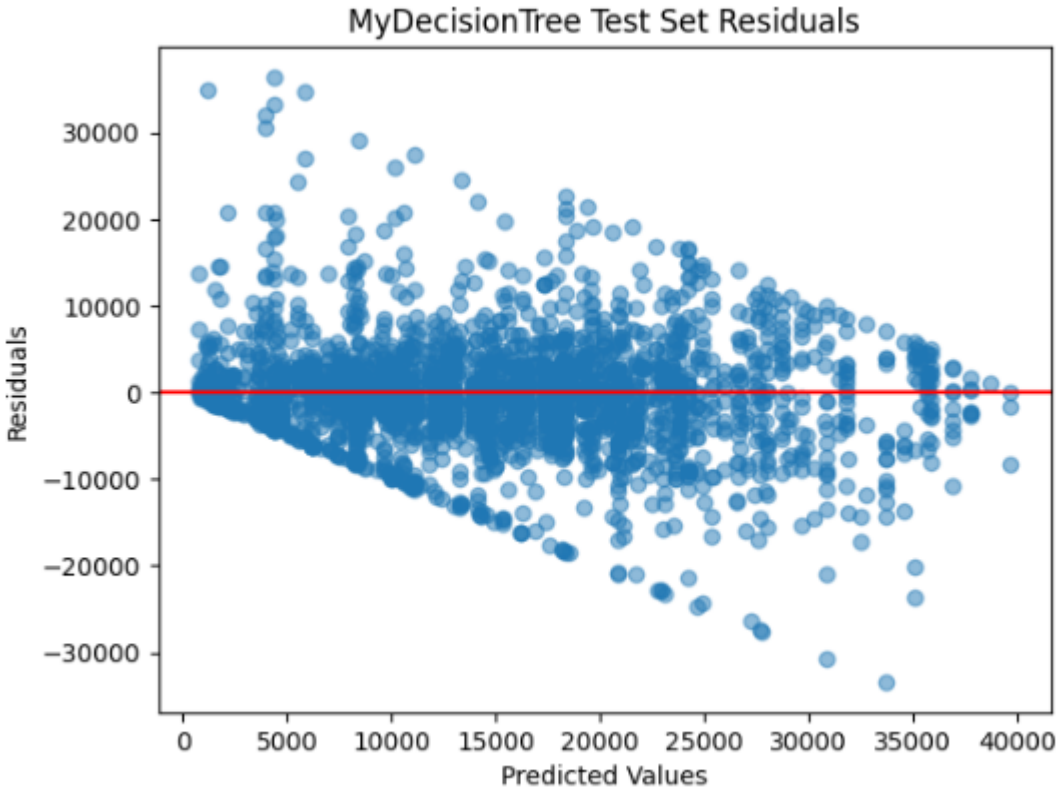
Rozloženie reziduál na trénovacej množine: reziduá sa nezdajú byť náhodne rozložené okolo nulovej čiary. Namiesto toho existuje vzor, kde sú reziduá pre nižšie predpovedané hodnoty bližšie k nule, a ako predpovedaná hodnota rastie, rozptyl reziduál sa zväčšuje, čo naznačuje, že presnosť predpovedí modelu je nižšia pri vyšších hodnotách.

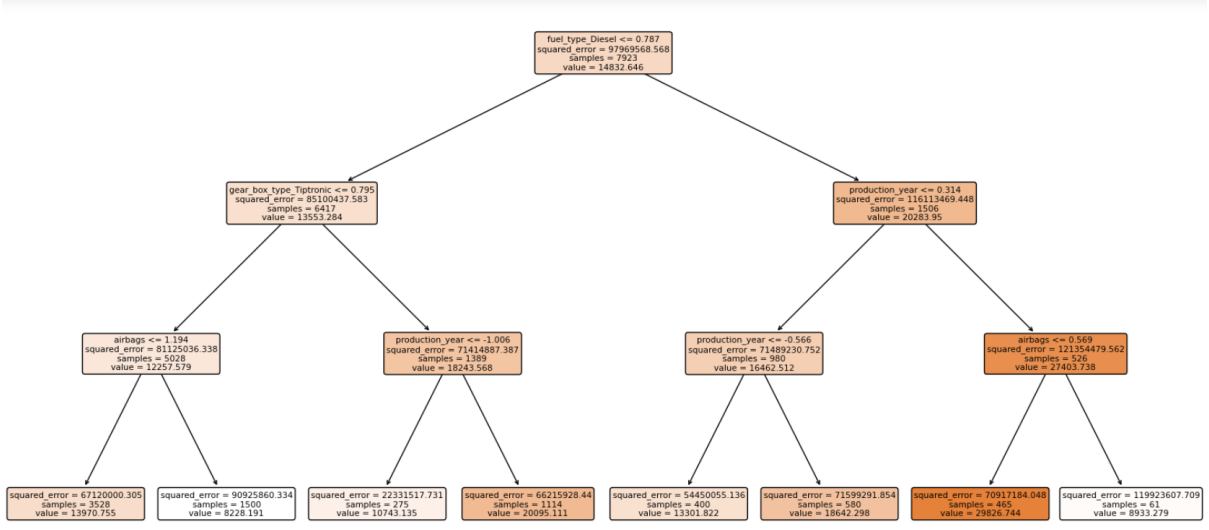
Trend reziduálov: Reziduá zobrazujú mierny „lievikovitý“ tvar s väčším rozptýlením, keď predpovedané hodnoty rastú, čo indikuje

možný trend v reziduáloch. Tento trend naznačuje, že model môže podceňovať skutočné hodnoty na nižšom konci predpovedaného rozsahu a nadhodnocovať na vyššom konci.

Rozloženie reziduál na testovacej množine: testovacia sada reziduál, preukazujú preukazuje podobný vzor ako trénovacia sada reziduál, ktorý naznačuje, že chyba predikcie modelu sa líši v závislosti od veľkosti predpovede, pričom väčšie chyby sa vyskytujú pri vyšších predpovedaných hodnotách.

Trend reziduálov: pozorovaný trend v reziduáloch oboch sád naznačuje, že model možno nedokáže dostatočne zachytiť zložité vzorce v dátach, obzvlášť pri vyšších hodnotách. To môže byť spôsobené nelineárnymi vzťahmi, ktoré rozhodovací strom danej hĺbky a zložitosti nezachytáva.


MyDecisionTree Test Set Residuals
Residuals
Predicted Values
30000
20000
10000
0
−10000
−20000
−30000
0
5000
10000
15000
20000
25000
30000
35000
40000


fuel_type_Diesel <= 0.787
squared_error = 97969568.568
samples = 7923
value = 14832.646
gear_box_type_Tiptronic <= 0.795
squared_error = 85100437.583
samples = 6417
value = 13553.284
production_year <= 0.314
squared_error = 116113469.448
samples = 1506
value = 20283.95
airbags <= 1.194
squared_error = 81125036.338
samples = 5028
value = 12257.579
production_year <= -1.006
squared_error = 71414887.387
samples = 1389
value = 18243.568
production_year <= -0.566
squared_error = 71489230.752
samples = 980
value = 16462.512
airbags <= 0.569
squared_error = 121354479.562
samples = 526
value = 27403.738
squared_error = 67120000.305
samples = 3528
value = 13970.755
squared_error = 90925860.334
samples = 1500
value = 8228.191
squared_error = 22331517.731
samples = 275
value = 10743.135
squared_error = 66215928.44
samples = 1114
value = 20095.111
squared_error = 54450055.136
samples = 400
value = 13301.822
squared_error = 71599291.854
samples = 580
value = 18642.298
squared_error = 70917184.048
samples = 465
value = 29826.744
squared_error = 119923607.709
samples = 61
value = 8933.279

Rozdelenia a features: strom rozdeľuje dáta na základe príznakov ako 'gear_box_type_Tiptronic', 'fuel_type_Diesel', 'production_year' a 'airbags'. Tieto príznaky model považuje za dôležité pri predpovedaní cieľovej premennej.

Squared errors: každý uzol na strome poskytuje metriku 'stvorec chyby'. Táto metrika kvantifikuje variabilitu cieľovej premennej v rámci uzla. Nižší stvorec chyby naznačuje, že predpovede uzla sú presnejšie.

Veľkosť Vzorky: každý uzol tiež zobrazuje počet 'vzoriek', ktoré spadajú do tej časti rozhodovacieho procesu. Ako sa posúvame po strome nadol, veľkosť vzorky v každom uzle zvyčajne klesá, pretože sú dáta ďalej rozdeľované.

Value: meranie centrálnej tendencie cieľovej premennej pre vzorky v uzle.

Listové uzly: koncové body stromu, nazývané listové uzly, reprezentujú konečné predpovede pre vzorky, ktoré ich dosiahnu.

- Trénovanie stromového súborového modelu

Natrénoval som náhodný les (random forest regressor) s použitím krížovej validácie (cross validation) s piatimi foldmi a desiatimi iteráciami, a s rôznymi hodnotami hyperparametrov. Najlepší model mal nasledujúce hyperparametre: 'n_estimators': 150, 'min_samples_split': 2, 'min_samples_leaf': 4, 'max_depth': 30. Najlepší model dosiahol na testovacej sade dát skóre R2 0.68. Na nasledujúcej tabuľke môžeme vidieť ďalšie výsledky najlepšieho modelu.

Dataset	R2 Score	MSE	RMSE
Train	0.83	16,650,678.73	4,080.52
Test	0.68	30,383,973.11	5,512.17

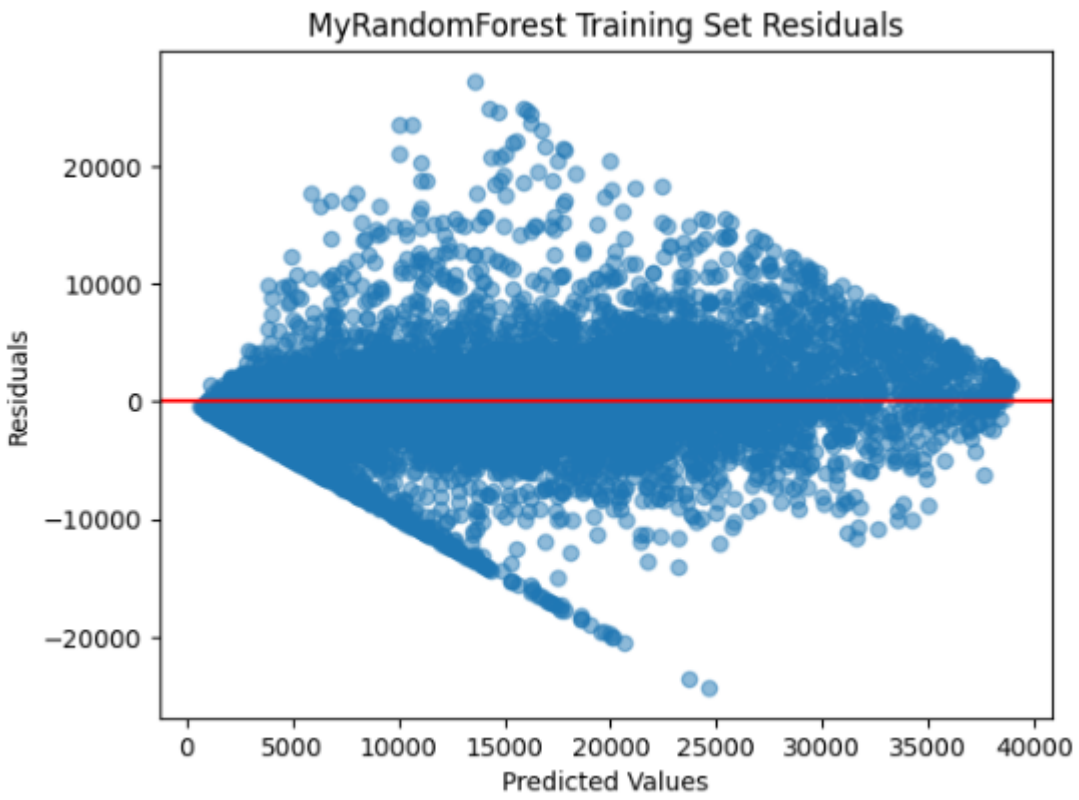
Po natrénovaní modelu sme vyhodnotili aj reziduály pre trénovaciu aj testovaciu množinu.

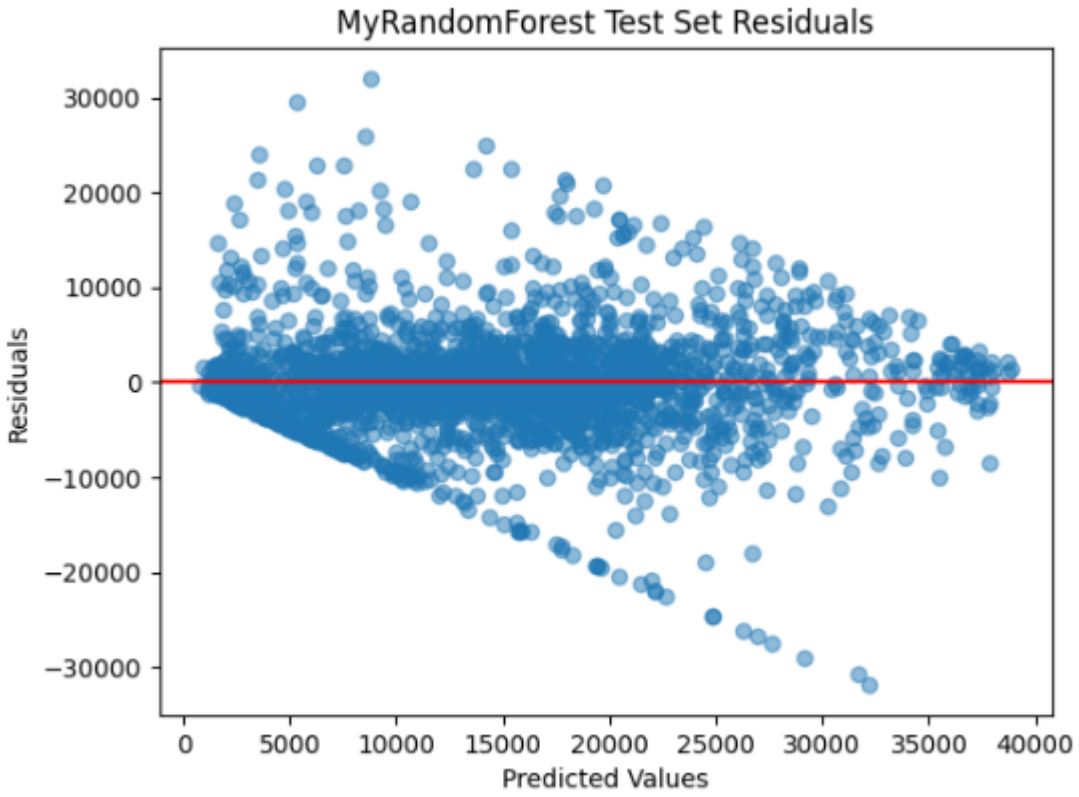
Rozloženie reziduál na trénovacej množine: reziduá sa rozširujú s rastúcou predpovedanou hodnotou. Rozptyl reziduál nie je konštantná v

celom rozsahu predpovedaných hodnôt. Predikcie modelu sú menej konzistentné pri vyšších hodnotách. Väčšina dát sa zdá byť sústredená pri nižších predpovedaných hodnotách, pri vyšších predpovedaných hodnotách sa reziduá stávajú rozptýlenejšími. To naznačuje, že model predpovedá nižšie hodnoty s vyššou presnosťou ako vyššie hodnoty.

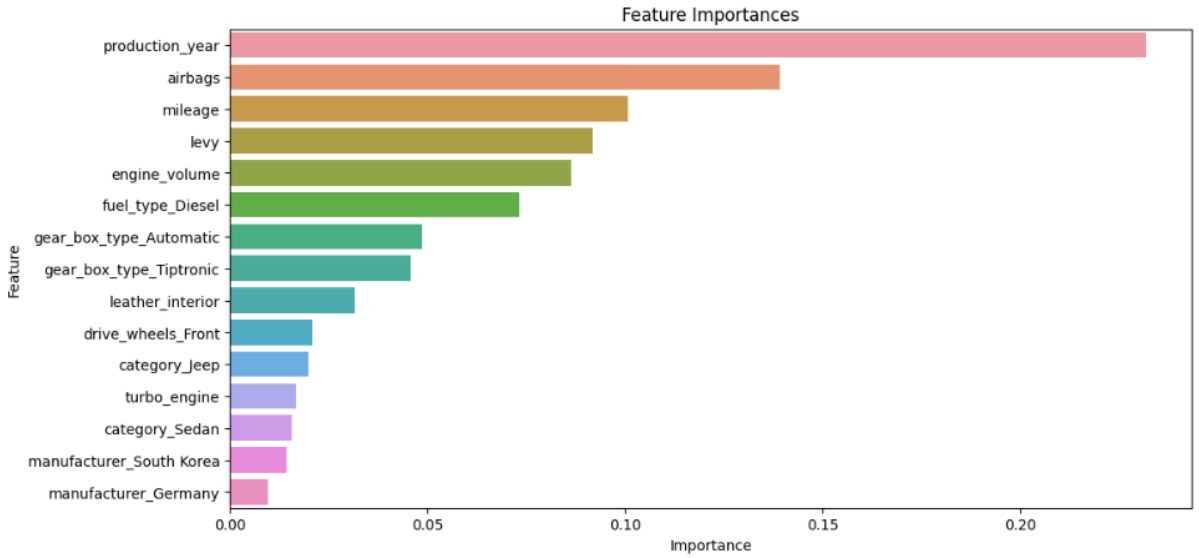
Trend reziduálov: zväčšujúce sa rozptýlenie reziduál pri vyšších predpovedaných hodnotách naznačuje, že predikcie modelu sa stávajú menej spoľahlivými s rastúcou hodnotou, čo môže indikovať obmedzenia modelu pri zachytávaní vzťahov v dátach pre tieto vyššie rozsahy.

Rozloženie reziduál a trend sú podobné na testovacej množine.





Na nasledujúcom obrázku môžeme vidieť dôležitosť prvkov modelu pre 15 najlepších prvkov.



- Model SVM

Trénoval som stroj SVM s použitím hyperparametrov C, gamma a kernel. Využil som nástroj GridSearchCV, ktorý preverí všetky možné kombinácie a identifikuje najvhodnejšie parametre pre model. Počet krížových validácií (cv) som nastavil na hodnotu tri. Najoptimálnejšie parametre, ktoré som identifikoval, boli: C rovné 100, gamma nastavená na 'scale' a typ jadra 'linear'. Najefektívnejší model dosiahol v testovacej sade dát skóre R2 vo výške 0.34. V nasledujúcej tabuľke sú prezentované ďalšie výsledky tohto najúčinnejšieho modelu:

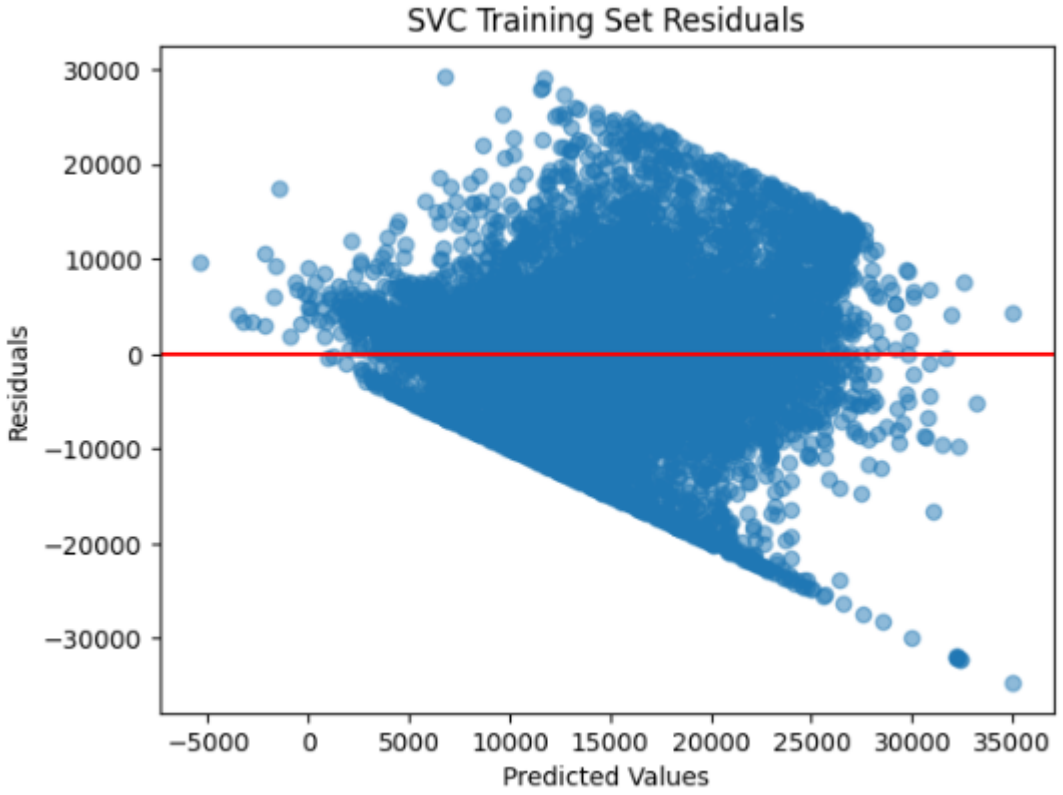
Dataset	R2 Score	MSE	RMSE
Train	0.34	64524352.34	8032.71
Test	0.34	63065151.62	7941.36

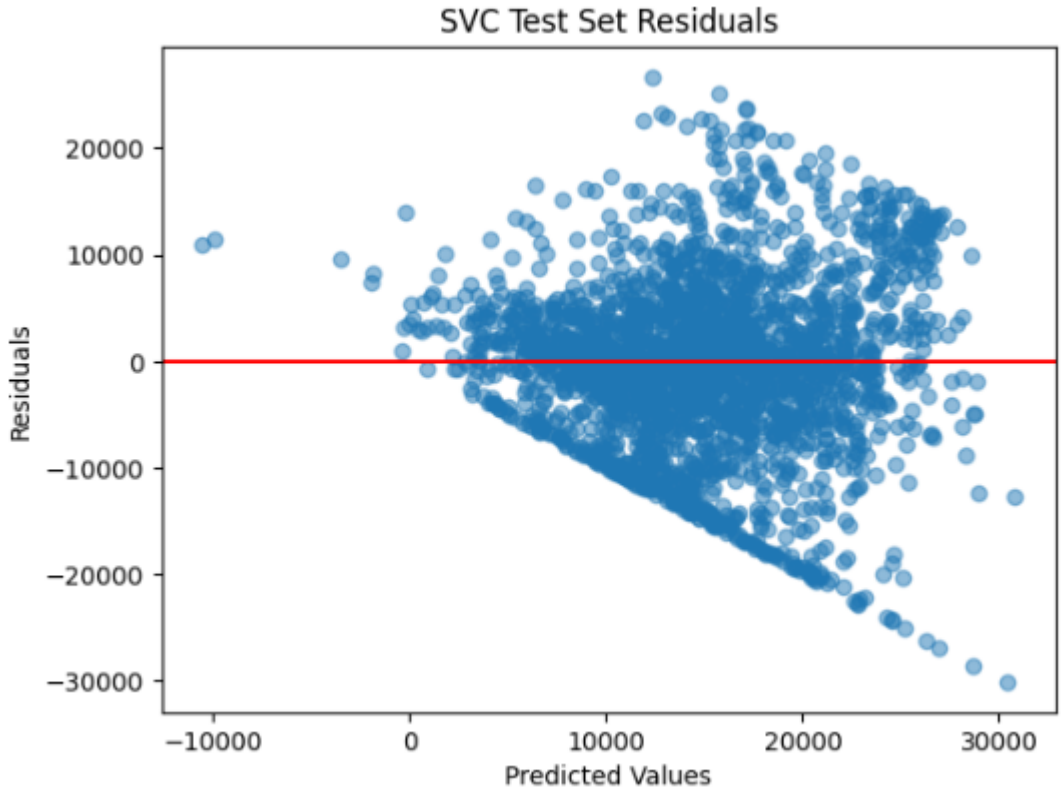
Po natrénovaní modelu sme vyhodnotili aj reziduály pre trénovaciu aj testovaciu množinu.

Rozloženie reziduál na trénovacej množine: reziduá nie sú symetricky rozložené okolo nulovej čiary (horizontálna červená čiara), čo naznačuje potenciálnu zaujatosť v predikcii. Rozptyl reziduál sa zdá byť väčší s rastúcimi predpovedanými hodnotami, čo naznačuje heteroskedasticitu, teda že variancia chýb nie je konštantná naprieč všetkými úrovňami nezávislých premenných.

Trend reziduálov: reziduá ukazujú určitý vzor alebo trend. Keď predpovedané hodnoty rastú, reziduá sa zdajú byť viac rozptýlené, a to nad aj pod nulovou čiarou. Tento vzor naznačuje, že model môže byť menej presný pri vyšších hodnotách. Ideálne by boli reziduá náhodne rozptýlené okolo horizontálnej osi bez rozpoznateľného vzoru, čo by naznačovalo, že model zachytil všetky relevantné vzorce v dátach.

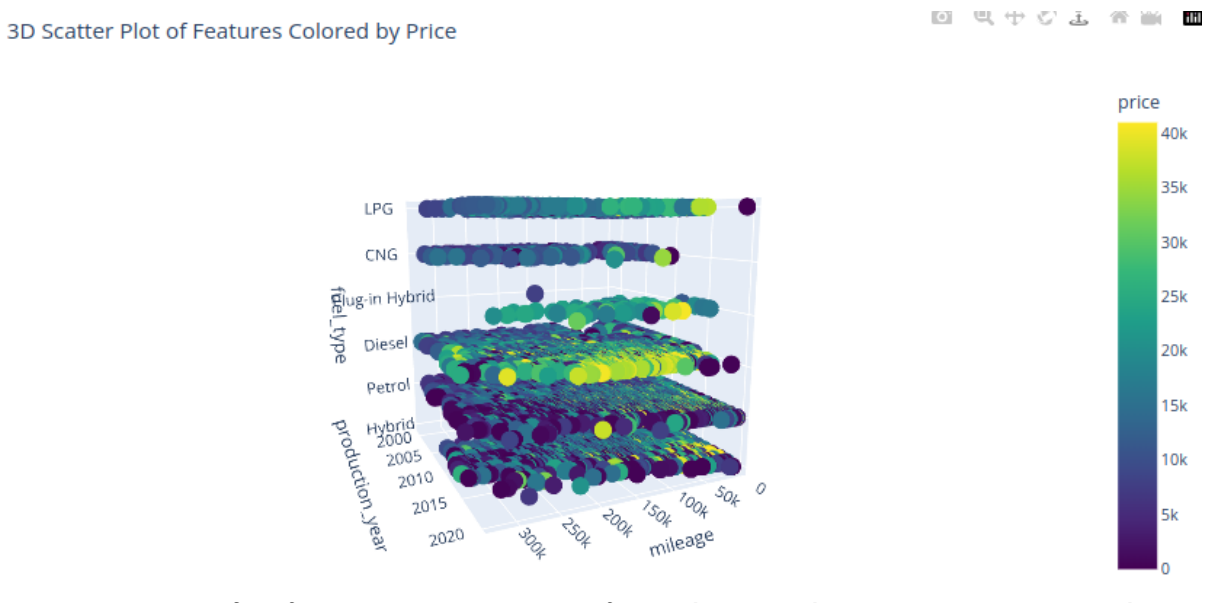
Rozloženie reziduál a trend sú podobné na testovacej množine.


SVC Training Set Residuals
Residuals
Predicted Values
30000
20000
10000
0
−10000
−20000
−30000
−5000
0
5000
10000
15000
20000
25000
30000
35000


SVC Test Set Residuals
Residuals
Predicted Values
20000
10000
0
−10000
−20000
−30000
−10000
0
10000
20000
30000

- Analýza troch príznakov

Snažil som sa vybrať charakteristiky tak, aby sme dosiahli čo najlepšiu vizualizáciu v 3D priestore. Po niekoľkých krokoch experimentovania som sa rozhodol pre charakteristiky: rok výroby, typ paliva a najazdené kilometre. Na nasledujúcom obrázku môžeme vidieť 3D graf, ktorý ilustruje tieto charakteristiky, pričom sú farebne odlíšené podľa charakteristiky ceny:



Z tejto vizualizácie by sme mohli vyvodiť určité vzorce, ako napríklad vzťah medzi vekom vozidla a jeho cenou, ako najazdené kilometre ovplyvňujú cenu, alebo či sú niektoré typy palív spojené s vozidlami s vyššou cenou. Napríklad by sme mohli pozorovať, že novšie autá (rok výroby bližšie k roku 2020) majú vyššiu cenu, čo je bežné očakávanie. Podobne, vozidlá s nižším počtom najazdených kilometrov by tiež mohli byť viditeľne drahšie. Môžeme tiež vidieť, že vozidlá typu Plug-in Hybrid sa začali predávať od roku 2010.

- Minimalizácia množinu parametrov na 3 dimenzie

Za účelom minimalizácie dimenzionality pomocou metódy PCA sme nastavili parameter n_components na hodnotu tri. Pri konštrukcii

grafu sme aplikovali farebné rozlíšenie bodov na základe monitorovaného parametra – ceny.

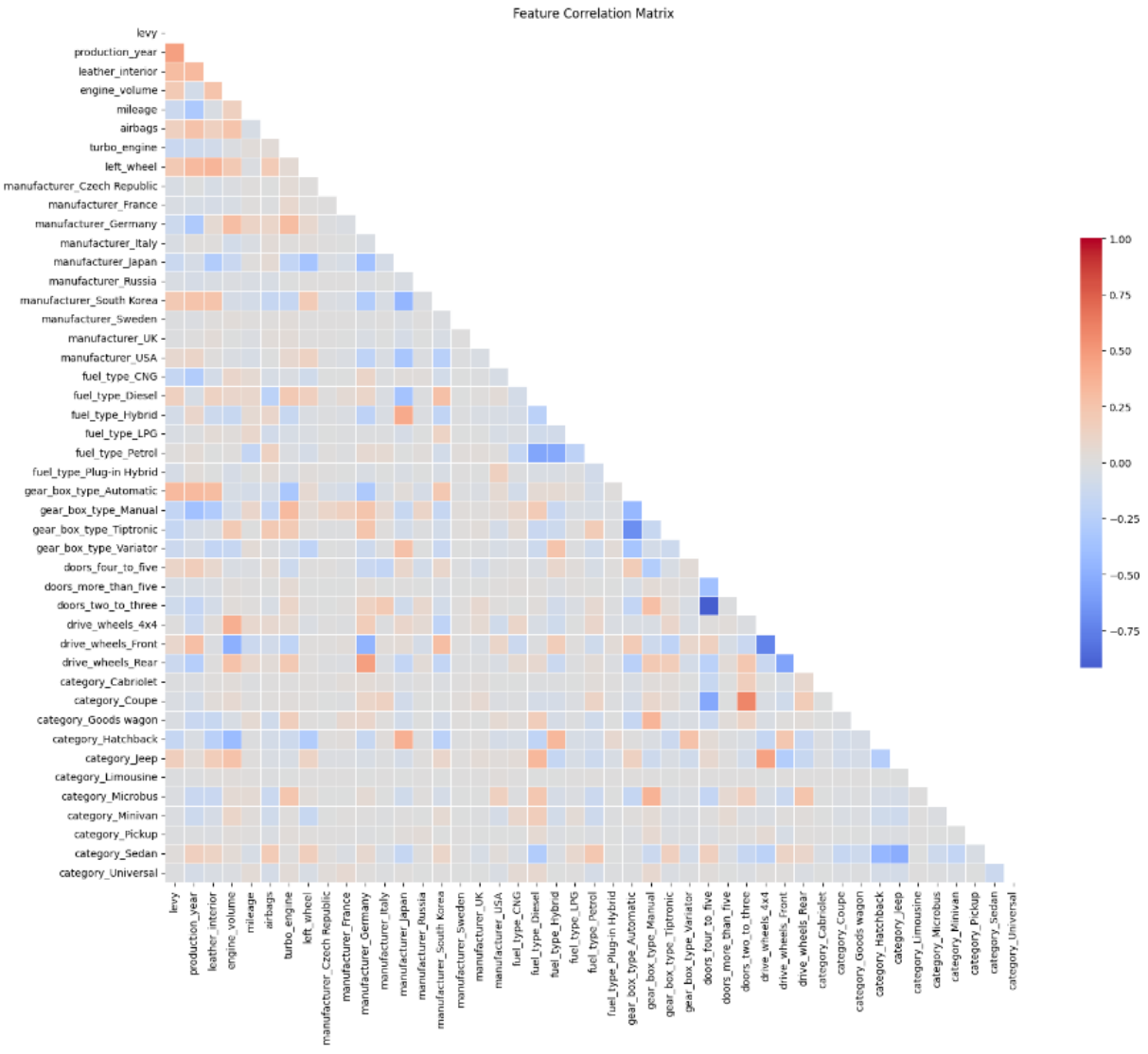
- Porovnanie grafov navzájom

Pri selekcii vhodných príznakov je jednoduchšie identifikovať vzťahy medzi nimi. Rozloženie bodov v grafe je pri tomto prístupe rozložené rovnomernejšie v porovnaní s metódou PCA. V procese minimalizácie prostredníctvom PCA nie sú vzťahy medzi charakteristikami jednoznačne viditeľné, pretože táto metóda zohľadňuje vzťahy v kontexte celého datasetu, na základe ktorých sú následne určené hodnoty pre jednotlivé dimenzie. Graf vytvorený metódou PCA má skôr tendenciu tvoriť zhluky než ukazovať rovnomerné rozloženie.

Natrénovanie najlepšieho modelu pre zmenšenú množinu

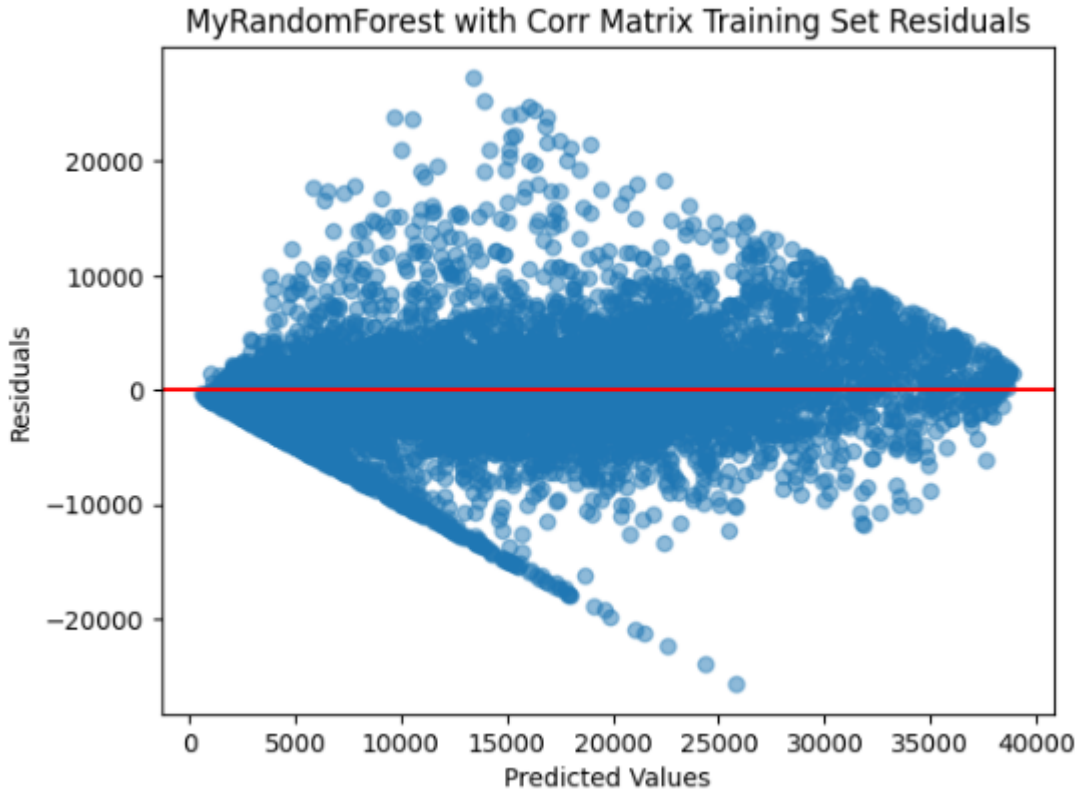
- podľa korelačnej matice:

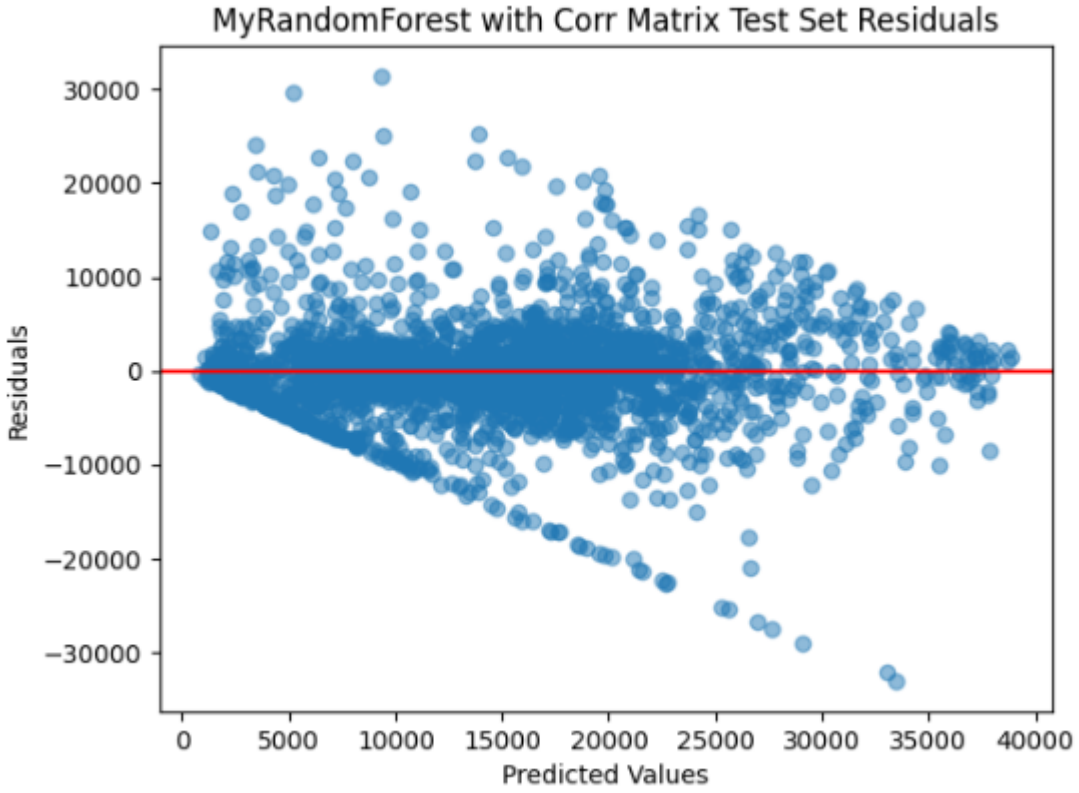
Pred začiatkom trénovania modelu bolo nevyhnutné vykonať analýzu dát prostredníctvom korelačnej matice. Preto sme vytvorili korelačnú maticu, ktorá nám poskytla hodnoty korelácie vo vzťahu k cieľovému parametru – cene. Pri analýze sme využili absolútne hodnoty korelácií, keďže aj záporné hodnoty indikujú existenciu negatívneho lineárneho vzťahu medzi charakteristikami. Stanovili sme parameter correlation_threshold, ktorý nám umožnil odfiltrovať tie charakteristiky, ktoré vykazovali koreláciu väčšiu ako 0.6 s cieľovým parametrom.



Pre trénovanie modelu sme zvolili metódu náhodných lesov, pretože sa ukázala byť najefektívnejšou. Opätovne sme využili nástroj GridSearchCV na optimalizáciu hyperparametrov. V nasledujúcej tabuľke sú prezentované výsledky z modelu, po tom, čo boli vstupné parametre filtrované na základe hodnôt z korelačnej matice:

Dataset	R2 Score	MSE	RMSE
Train	0.83	16549626.46	4068.12
Test	0.68	30296315.26	5504.21



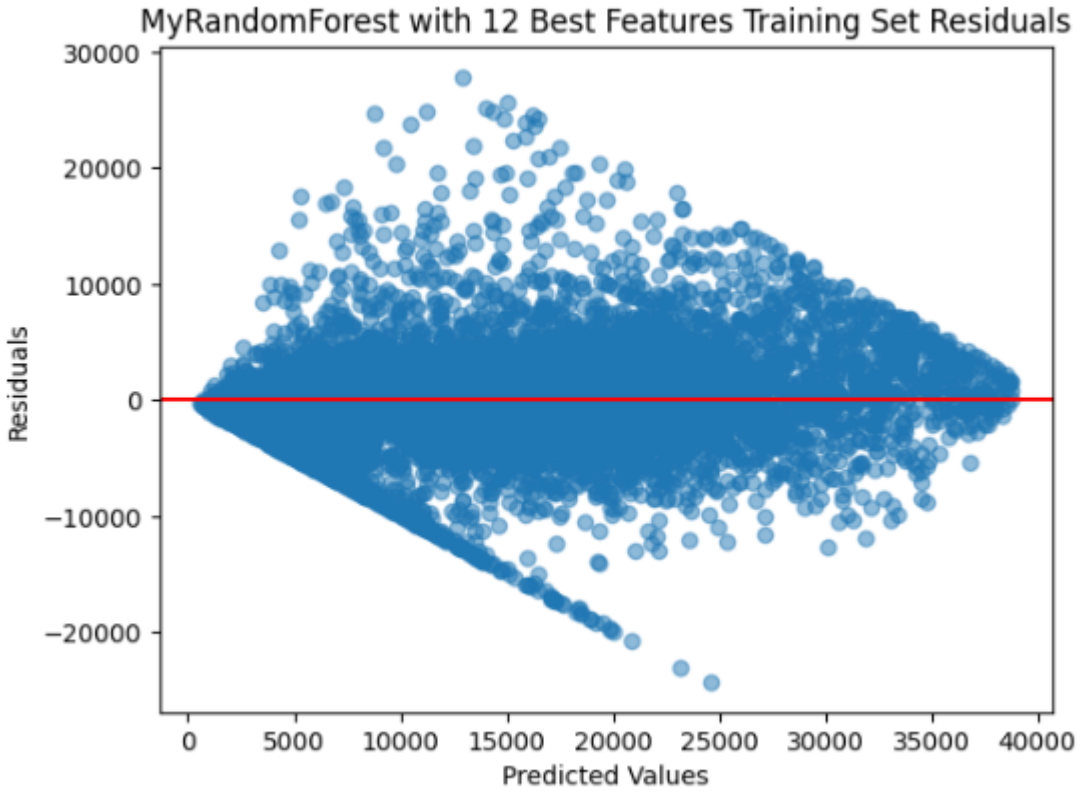


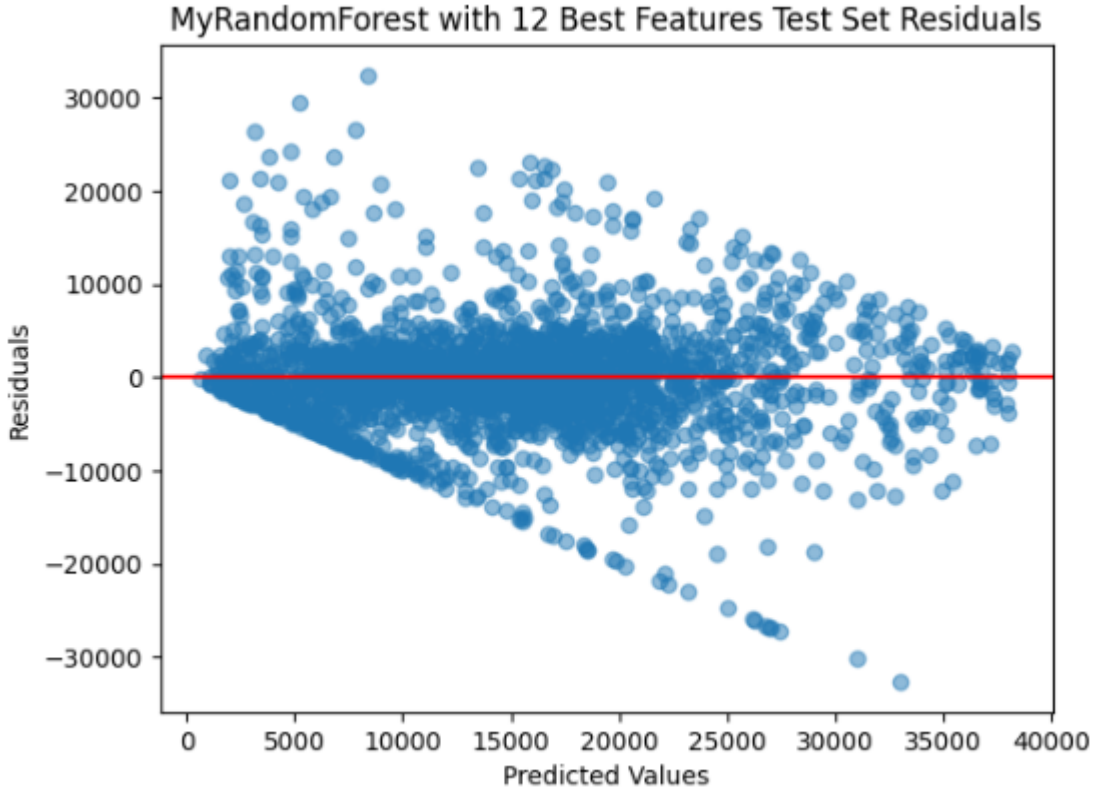
- podľa dôležitosti príznakov súborového modelu

Pri príprave trénovania modelu sme postupovali systematicky a zohľadnili sme dôležitosť príznakov určených modelom náhodných lesov. Na základe tohto hodnotenia sme vybrali dvanásť najrelevantnejších príznakov, ktoré zahŕňajú: 'production_year', 'airbags', 'mileage', 'levy', 'fuel_type_Diesel', 'engine_volume', 'gear_box_type_Tiptronic', 'leather_interior', 'drive_wheels_Front', 'category_Jeep', 'turbo_engine', 'gear_box_type_Automatic'.

Výsledky trénovania sú nasledovné:

Dataset	R2 Score	MSE	RMSE
Train	0.81	18120734.60	4256.85
Test	0.65	33240446.26	5765.45


MyRandomForest with 12 Best Features Training Set Residuals
Residuals
Predicted Values
30000
20000
10000
0
−10000
−20000
0
5000
10000
15000
20000
25000
30000
35000
40000

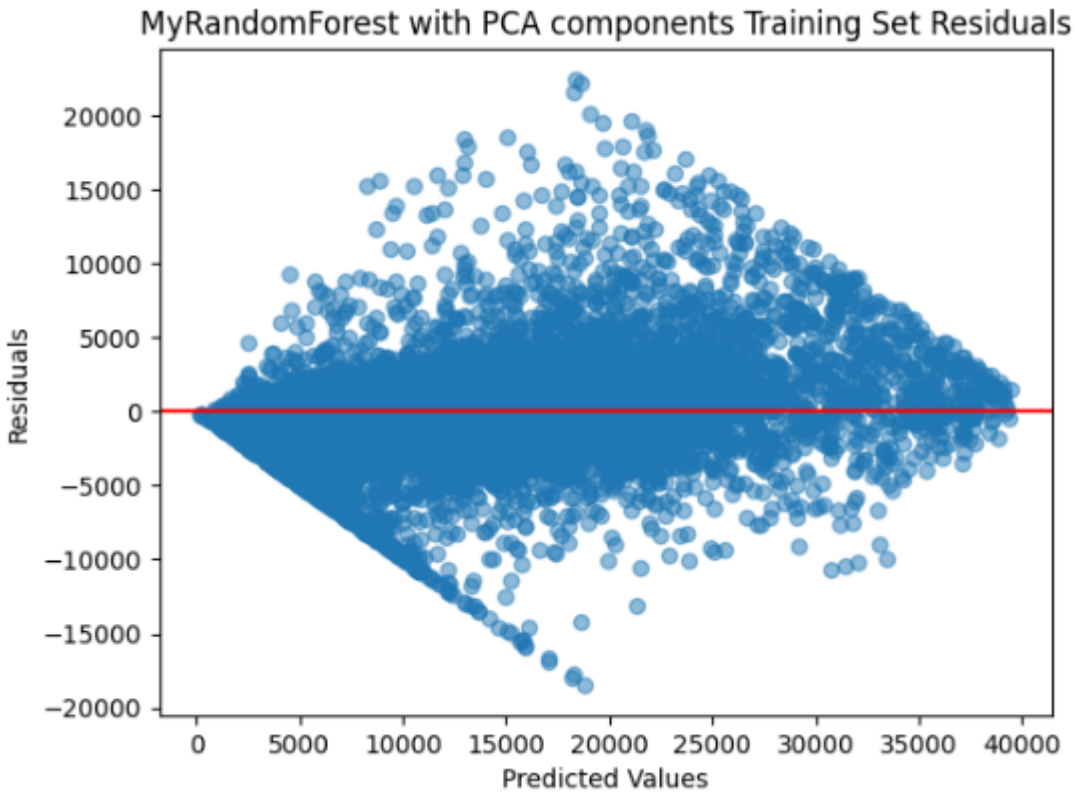

MyRandomForest with 12 Best Features Test Set Residuals
Residuals
Predicted Values
30000
20000
10000
0
−10000
−20000
−30000
0
5000
10000
15000
20000
25000
30000
35000
40000

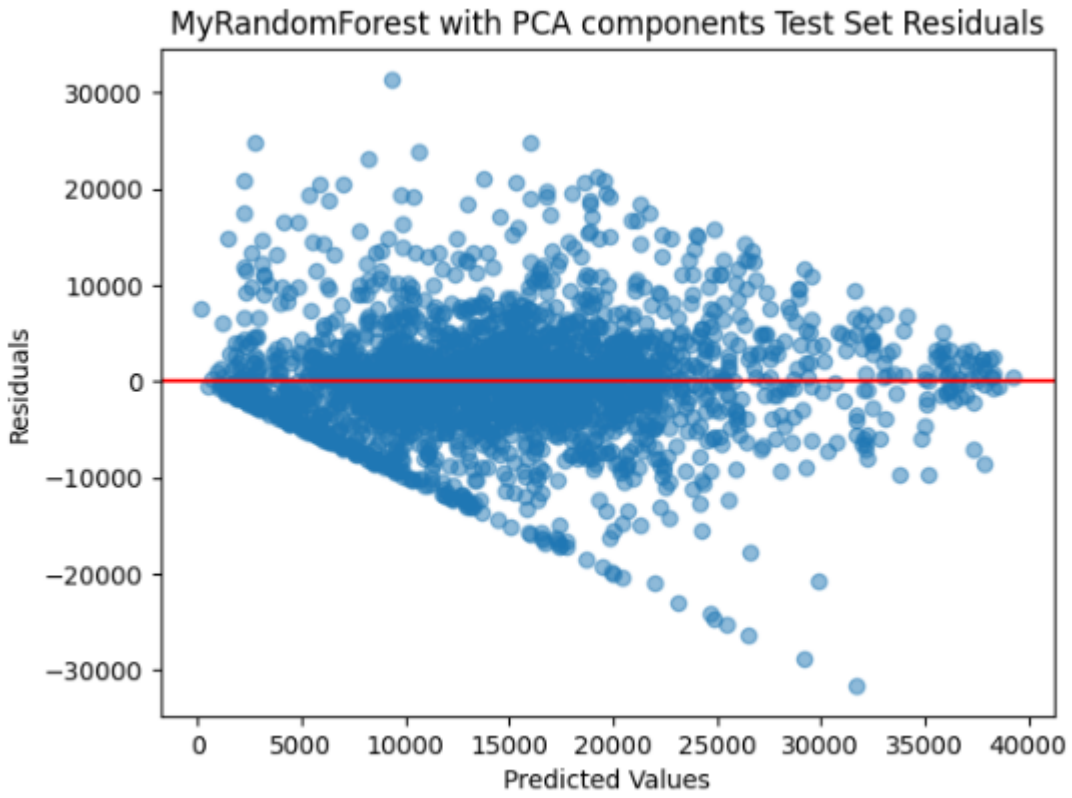
- podľa variancie pomocou PCA

Pri príprave modelu s použitím metódy hlavných komponentov (PCA) sme namiesto výberu počtu charakteristík určili hodnotu zachovanej variancie, ktorá vyjadruje percentuálny podiel zachovanej variancie po redukcii dimenzií. Tento ukazovateľ sme nastavili na úroveň 0,9. Po uplatnení daného prahu nám z pôvodného súboru údajov zostalo 30 hlavných komponentov.

Výsledky trénovania sú nasledovné:

Dataset	R2 Score	MSE	RMSE
Train	0.86	13188804.59	3631.64
Test	0.63	35131202.20	5927.16





- Porovnanie výsledkov

Najefektívnejšími modelmi v rámci tejto časti úlohy sa ukázali byť modely trénované podľa korelačnej matice a samotný model náhodných lesov. Ostatné modely, s výnimkou SVM, vykázali porovnateľné výsledky. Model SVM, na druhej strane, preukázal najnižšiu úroveň výkonnosti.

Pokiaľ ide o reziduá modelov, každý z nich vykazoval podobný trend v ich rozložení.

Model	Train R2 Score	Train MSE	Train RMSE	Test R2 Score	Test MSE	Test RMSE
Decision Tree	0.65	34,123,009.31	5,841.49	0.61	37,321,020.64	6,109.09
Random Forest	0.83	16,650,678.73	4,080.52	0.68	30,383,973.11	5,512.17
SVM	0.34	64,524,352.34	8,032.71	0.34	63,065,151.62	7,941.36
RF Corr Matrix	0.83	16,549,626.46	4,068.12	0.68	30,296,315.26	5,504.21
RF Best Features	0.81	18,120,734.60	4,256.85	0.65	33,240,446.26	5,765.45
RF via PCA components	0.86	13,188,804.59	3,631.64	0.63	35,131,202.20	5,927.16

Bonus:

- Zhlukovanie dát

Pri určovaní počtu zhlukov som využil metódu Elbow v rozmedzí od 1 do 10 zhlukov. Na obrázku je zreteľné, že "inertia" - bod, kde je pokles variance v rámci zhlukov pomalší, je možné identifikovať pri troch zhlukoch.