# Výsledná správa k semestrálnemu projektu

Bc. Miroslav Čulík a Bc. Andrej Gáfrik

Fakulta informatiky a informačných technológií,
Slovenská technická univerzita v Bratislave,
Ilkovičova 2, 842 16 Bratislava
info@fiit.stuba.sk
https://www.fiit.stuba.sk

**Abstract.** V tomto semestrálnom projekte sme sa zamerali na úlohu predikcie cien nehnuteľností v americkej oblasti King County podľa dostupných atribútov. Dáta sme najprv vo fáze exploratívnej analýzy dôkladne preskúmali a navrhli sme riešenia, ktoré sme realizovali vo fáze predspracovania dát, zároveň sme vytvorili 2 nové atribúty, konkrétne recon_age a price_per_sqft, ktoré sú kombináciou pôvodných atribútov. Okrem toho sme experimentálne vyskúšali niekoľko modelov predikujúcich závislú premennú ceny nehnuteľnosti. Tieto modely sme vyhodnotili zvolenými metrikami. V tretej fáze sme pridali ďalšie modely, pričom sme na nich vykonali ladenie hyperparametrov. Nakoniec sme vybrali model Náhodného lesa, ktorý sústavne dosahoval najlepšie výsledky podľa metriky RMSE a použili sme ho na predikciu cien testovacích dát. Všetky výsledky sú podrobne zdokumentované v práci, prípadne v priložených súboroch.

## 1 Opis problému a dát

V našom projekte by sme chceli sa chceli zaoberať problémom predikcie cien nehnuteľností na základe dostupných atribútov o týchto nehnuteľnostiach. Táto predikčná úloha môže pomôcť lepšie odhadnúť skutočnú cenu nehnuteľností pri ich predajoch a nákupoch.

Nami zvolený dataset obsahuje údaje o domoch, ktoré boli predané v USA (oblasť King County v štáte Washington) v období máj 2014 až máj 2015. Dáta sme získali z portálu kaggle.com vo formáte **.csv** súboru[1].

Po vykonaní exploratívnej analýzy bolo zistené, že naše dáta majú celkovo 21 atribútov (stĺpcov), z toho je 20 atribútov numerických a jeden atribút obsahuje dátumové hodnoty. Spolu náš dataset obsahuje 21613 záznamov, pričom chýbajúce hodnoty v ňom nie sú prítomné.

### 1.1 Charakteristika dát

Ako bolo povedané vyššie, náš dataset má 21 stĺpcov, pričom okrem jedného prípadu ide o numerické atribúty. Charakteristika jednotlivých atribútov:

[1] https://www.kaggle.com/harlfoxem/housesalesprediction

- **id** - index inštancie
- **date** - dátum predaja nehnuteľnosti
- **price** - cena nehnuteľnosti (pravdepodobne v USD)
- **bedrooms** - počet spální v nehnuteľnosti
- **bathrooms** - počet kúpeľní v nehnuteľnosti
- **sqft_living** - rozloha obytného priestoru v stopách štvorcových (square feet)
- **sqft_lot** - rozloha celého pozemku v stopách štvorcových (square feet)
- **floors** - počet poschodí
- **waterfront** - či z nehnuteľnosti je výhľad na pobrežie
- **view** - index kvality výhľadu z nehnuteľnosti
- **condition** - stav nehnuteľnosti
- **grade** - stupeň vnútorného zariadenia
- **sqft_above** - rozloha obytného priestoru v nadzemnej časti budovy v stopách štvorcových (square feet)
- **sqft_basement** - rozloha obytného priestoru suterénu v stopách štvorcových (square feet)
- **yr_built** - rok postavenia nehnuteľnosti
- **yr_renovated** - rok renovovania nehnuteľnosti
- **zipcode** - poštové smerovacie číslo
- **lat** - zemepisná šírka (severná šírka)
- **long** - zemepisná dĺžka (západná dĺžka)
- **sqft_living15** - rozloha obytného priestoru pre najbližších 15 susedných nehnuteľností
- **sqft_lot15** - rozloha celého pozemku pre najbližších 15 susedných pozemkov

## 2 Stručný opis prác iných autorov

V rámci vyhľadávania vedeckých a nevedeckých článkov, ktoré pracovali s podobnými dátami a riešili podobnú úlohu ako my, sme vybrali nasledujúce práce, ktoré nás inšpirovali vo výbere prístupov strojového učenia ako aj metrík na vyhodnotenie týchto prístupov.

### 2.1 Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia[1]

Autori tohto článku [1] sa zamerali na predikciu cien nehnuteľností v austrálskom veľkomeste Melbourne. Výsledné modely boli trénované na korpuse vyše 20 000 záznamov. Autori k outlierom pristupovali individuálne podľa stĺpca, v ktorom sa nachádzali (viac informácií však v článku neposkytli). Predikovaný stĺpec s cenou nehnuteľností bol transformovaný pomocou logaritmu. Pri redukcii dimenzionality boli použité techniky **Stepwise** a **Boosting** na získanie najsilnejších prediktorov, pre model SVM bola taktiež použitá technika PCA. Všetky výsledky boli porovnávané s východzím modelom lineárnej regresie. Najlepšie výsledky boli dosiahnuté pri SVM, avšak veľké rozdiely medzi train MSE a eval MSE naznačujú preučený model.

| Model | Train MSE | Eval. MSE | Eval. Ratio |
|---|---|---|---|
| Lineárna regresia | 0.0948 | 0.0994 | **1.00** |
| Polynomiálna regresia | 0.0773 | 0.0832 | **0.84** |
| Regresný rozhodovací strom | 0.0925 | 0.0985 | **0.99** |
| Neurónová sieť | 0.2657 | 0.2749 | **2.77** |
| Stepwise + SVM | 0.0558 | 0.0615 | **0.62** |
| Stepwise + vyladený SVM | 0.0480 | 0.0561 | **0.56** |
| PCA + SVM | 0.0721 | 0.0810 | **0.82** |
| PCA + vyladený SVM | 0.0474 | 0.0728 | **0.74** |

**Table 1.** Výsledky predikcií pri použití rôznych modelov[1]

### 2.2 Research on the Sustainable Development of Urban Housing Price Based on Transport Accessibility:A Case Study of Xi'an, China [2]

Autori [2] v článku skúmajú vplyv prístupnosti verejnej dopravy na predikciu reálnych cien nehnuteľností. Tento článok pracuje s dátami obsahujúcimi nehnuteľnosti a možnosti transportu(cestná sieť, autobusové linky a linky metra) v hlavnej mestskej časti čínskeho mesta Xi'an. Použitím tradičného hedonického cenového modelu (z angl. *traditional hedonic price model*) a algoritmom náhodného lesa (z angl. *Random Forest*) autori porovnávajú situácie s prítomným, resp. neprítomným atribútom znázorňujúcim prístupnosť verejnej dopravy(prístupnosť autobusovej verejnej dopravy a metra). Výsledkom tejto práce bolo zistenie, že schopnosť modelov predikovať reálnu cenu nehnuteľností po pridaní atribútu prístupnosti verejnej dopravy je štatisticky zlepšená oproti situácii, kedy atribút prístupnosti verejnej dopravy prítomný nie je.

### 2.3 Regression using sklearn on KC Housing Dataset [3]

Tento nevedecký článok [3] sa zaoberal predikciou cien nehnuteľností na našom datasete. V článku sa autor šikovne vyhol predspracovaniu dát, avšak dokázal pri jednoduchších modeloch dosiahnuť zaujímavé výsledky.

| Model | Eval. RMSE | Train $R^2$ | Eval. $R^2$ |
|---|---|---|---|
| Jednoduchá lineárna regresia (sqft_living) | 254,289.15 | 0.492 | 0.496 |
| Jednoduchá lineárna regresia (grade) | 263,387.61 | 0.442 | 0.46 |
| Viacnásobná lineárna regresia (features1) | 239,014.4 | 0.548 | 0.555 |
| Viacnásobná lineárna regresia (features2) | 205,244.56 | 0.655 | 0.672 |
| Polynomiálna regresia(features1) - stupeň 2 | 175,814.11 | 0.758 | 0.759 |
| Polynomiálna regresia(features1) - stupeň 3 | 207,567.72 | 0.776 | 0.664 |

**Table 2.** Výsledky predikcií pri použití rôznych modelov[3]

## 3 Predspracovanie a výber atribútov

V našom datasete sme detekovali niekoľko stĺpcov, ktoré obsahovali vychýlené hodnoty (z angl. *outliers*). Na detekciu sme používali metódu tzv. fúzov (z angl. *whiskers*), ktorá považovala za outlier hodnoty mimo intervalu (25 percentil - 1,5*IQR; 75 percentil + 1,5*IQR).

Outlierov sme nahrádzali 3 rôznymi spôsobmi:

- **5-95 percentil** - hodnoty nachádzajúce sa pod ľavým "fúzom" boli nahradené 5 percentilom stĺpca, hodnoty nad pravým "fúzom" naopak 95 percentilom, takto boli nahradené vychýlené hodnoty v stĺpcoch price, sqft_living, sqft_above, sqft_living15
- **medián** - všetci outlieri boli nahradení mediánom daného stľpca, napr. pri sqft_lot, sqft_lot15
- **priemer** - analogický postup, ako pri mediáne, len nahrádzanie priemernou hodnotou stĺpca, napr. sqft_basement

Po podrobnej analýze datasetu sme prišli k záveru, že vo väčšine prípadov sú outlieri reálne hodnoty a nie chyby (kontrola napr. výpisom rozlohy najdrahších nehnuteľností, počtu izieb, poschodí...). Preto sme sa rozhodli na základe odhadu z grafu 1, ktorý indikoval rastúcu chybu pri zvyšujúcej sa cene nehnuteľnosti, brať do úvahy tie inštancie, ktoré boli nižšia ako cena 815 000 USD. V stĺpcoch bathroom a bedrooms sme identifikovali záznamy, pri ktorých sme museli konštatovať, že vznikli chybou. Ide napr. o situáciu, kedy sme identifikovali záznam s počtom izieb 33. Keďže rozloha obytnej časti neodpovedala počtu izieb, rozhodli sme sa tento záznam vyhlásiť za chybu a hodnotu 33 nahradiť číslom 3. Na opačnej strane sme identifikovali 7 prípadov, pri ktorých bol počet izieb, ako aj kúpeľní 0. Keďže tieto záznamy mali aj podozrivo malú rozlohu, boli sme nútení pristúpiť k ich odstráneniu.

### 3.1 Nahradenie vychýlených hodnôt

Navrhli a zrealizovali sme nasledovný spôsob nahradenia identifikovaných vychýlených hodnôt:

- **price** - transformácia pomocou boxcox transformácie (kvôli pôvodnej nesymetrickosti dát) a nahradenie outlierov 5-95 percentilom (kvôli veľkému počtu odľahlých hodnôt na chvostoch rozdelenia)
- **bedrooms** - manuálna úprava inštancie s hodnotou 33 na 3
- **sqft_living** - transformácia pomocou logaritmu (pri pôvodnej nesymetrickosti dát nám logaritmus dokázal vyprodukovať najviac symetrické dáta) a nahradenie outlierov 5-95 percentilom (kvôli veľkému počtu odľahlých hodnôt na chvostoch rozdelenia)
- **sqft_lot** - transformácia pomocou boxcox transformácie (kvôli pôvodnej nesymetrickosti dát) a nahradenie outlierov mediánom (aj po boxcox transformácii zostalo na chvostoch rozdelenia veľké množstvo hodnôt, takže radšej sme pristúpili k nahrádzaniu strednou hodnotou)

- **sqft_above** - transformácia pomocou boxcox transformácie (kvôli pôvodnej nesymetrickosti dát) a nahradenie outlierov 5-95 percentilom (transformované rozdelenie hodnôt udávalo vhodný koeficient symetrie)
- **sqft_basement** - transformácia pomocou odmocniny (veľký výskyt nehnuteľností s hodnotou 0) a nahradenie outlierov priemerom (vyšší než medián, čo bola vhodná alternatíva pre outlierov sprava)
- **sqft_living15** - transformácia pomocou logaritmu(pri pôvodnej nesymetrickosti dát nám logaritmus dokázal vyprodukovať najviac symetrické dáta) a nahradenie outlierov 5-95 percentilom (transformované rozdelenie hodnôt udávalo vhodný koeficient symetrie)
- **sqft_lot15** - transformácia pomocou boxcox transformácie (kvôli pôvodnej nesymetrickosti dát) a nahradenie outlierov mediánom (veľký počet hodnôt na chvostoch rozdelenia aj po transformácii)

Pri stĺpci **zipcode** sme pristúpili k zakódovaniu jeho hodnôt binárnym kódovaním (binary_encoding), keďže sa jednalo o nominálny kategorický atribút.

### 3.2 Vytvorenie nových atribútov

Pristúpili sme aj k vytvoreniu nových atribútov:

- **recon_age** - popisuje počet rokov od poslednej rekonštrukcie nehnuteľnosti po rok predaja. Ak nebola uskutočnená žiadna rekonštrukcia, ide o rozdiel roku predaja nehnuteľnosti a roku postavenia.
- **price_per_sqft** - popisuje strednú hodnotu ceny na 1 stopu štvorcovú (square feet2) ako pomer ceny nehnuteľnosti a súčet atribútu sqft_living a sqft_basement podľa danej lokality, ktorá je reprezentovaná pôvodnými hodnotami atribútu zipcode

Napokon sme podľa atribútu **date** zoradili všetky inštancie a rozdelili sme ich na trénovaciu, validačnú a testovaciu množinu v pomere **70:20:10**. Keďže nám už atribút **date** nebol viac potrebný, zmazali sme ho.

### 3.3 Výber atribútov

Pri viacerých prístupoch sme museli riešiť problém výberu vhodných atribútov (*feature selection*). Implementovali sme nasledovné algoritmy na výber atribútov:

- *filter* - pomocou hodnoty korelácie s predikovaným atribútom
- *wrapper* - **SFS** (Sequential Feature Selection)
- *wrapper* - **RFE** (Recursive Forward Elimination)

## 4 Opis metód a ich vyhodnotenia

Pri výbere prístupov strojového učenia sme predovšetkým dbali na príbuzné práce, ako aj odporúčania z prednášok z predmetu Objavovanie znalostí. Na riešenie problému predikcie ceny nehnuteľností sme v nami prevedených experimentoch použili nasledovné prístupy strojového učenia:

- **Jednoduchá lineárna regresia** - pomocou jednej nezávislej premennej chceme predikovať závislú premennú ceny
- **Viacnásobná lineárna regresia** - pomocou viacerých nezávislých premenných chceme predikovať závislú premennú ceny, pričom každá nezávislá premenná je rádu 1
- **Polynomiálna regresia** - pomocou viacerých nezávislých premenných rádu vyššieho ako 1 predikujeme závislú premennú ceny. Počas našich experimentov vykonaných v tejto fáze sme skúsili polynomiálnu regresiu 2. a 3. stupňa.
- **Regresný rozhodovací strom** - Ide o upravenie rozhodovacích stromov, ktoré sa primárne používajú v úlohách klasifikácie. Regresné stromy pracujú na princípe učenia sa lokálnych lineárnych regresií pomocou aproximácie sínusovej funkcie s prihliadnutím na posun v pozorovaniach. Algoritmus si pri veľkom počte atribútov dokáže sám určiť, podľa ktorých bude modelovať závislú premennú (v našom prípade atribút **price**)
- **Náhodný les** - Vytvorí určený počet regresných rozhodovacích stromov, ktoré budú vytvorené z vybranej podmnožiny vstupného datasetu. Každý vytvorený regresný rozhodovací strom určuje podľa svojej stavby odhad predikovanej premennej, čoho výsledkom je priemerný odhad pre všetky stromy. Podobne ako pri regresných rozhodovacích stromoch ide o modifikáciu prístupu, ktorý pôvodne slúži v úlohách klasifikácie.

### 4.1 Metriky na vyhodnotenia metód

Na vyhodnotenie vyššie opísaných modelov sme sa rozhodli použiť tieto metriky:

1. **Mean squared error (MSE)** je základná metrika, ktorá zohľadňuje absolútny rozdiel medzi predikovanými hodnotami a skutočnými hodnotami. Vypočíta sa ako priemer rozdielov predikovaných a reálnych hodnôt umocnených na druhú. Táto metrika preferuje modely uprednostňujúce väčšinové výsledky, jej použitie môže viesť k zavádzajúcim alebo nepresným záverom, preto je ideálne ju používať v kombinácii s inými metrikami. Vzťah na výpočet MSE je nasledovný:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2$$

2. **Root mean squared error (RMSE)** je metrika odvodená od MSE. Ako už z názvu vyplýva, vypočíta sa odmocnením výsledku MSE. Nasleduje vzťah výpočtu.

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

3. **Root Mean Squared Log Error (RMSLE)** sa používa na výpočet pomeru predikovaných a reálnych hodnôt. Jej najlepšie využitie je v prípade, kedy pracujeme s veľmi veľkými hodnotami, teda aj MSE a RMSE narastajú

do veľkých rozmerov. RMSLE taktiež viac penalizuje menšie predikcie, ako väčšie pri rovnakej vzdialenosti od reálnej hodnoty.

$$RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(log(y_i) - log(\widehat{y_i}))^2}$$

4. $\mathbf{R}^2$ je štatistická metrika, ktorá znázorňuje mieru kvality regresného modelu. Vyjadruje, aký podiel variability závislej premennej model vysvetľuje. Nadobúda hodnoty od 0 po 1, pričom model s hodnotou 1 dokonalo predikuje hodnoty závislej premennej, pri hodnote 0 hovoríme o neužitočnom modeli. Vypočítame ju nasledovne:

$$R^2 = \frac{\sum(\widehat{y_i} - \overline{y_i})^2}{\sum(y_i - \overline{y_i})^2}$$

## 5 Experimenty

Po predspracovaní dát nami vytvorenou pipeline, sme mali dáta rozdelené v pomere 70:20:10 (trénovacia : validačná : testovacia množina). S takto pripravenými dátami sme začali trénovanie s vyššie opísanými modelmi. Najprv sme každý model natrénovali na trénovacej množine a zvalidovali na validačnej množine. Modely sme pre obe množiny s použitím opísaných metrík aj vyhodnotili. Okrem popísaných modelov sme vytvorili primitívny prediktor, ktorý pre všetky inštancie predikoval priemernú cenu inštancií predspracovanej trénovacej množiny (436,853 USD). Tento primitívny prediktor nám poslúžil ako východzí bod, podľa ktorého sme vedeli porovnávať všetky ostatné modely pre metriky MSE, RMSE a RMSLE. Čo sa týka použitých metrík, tak najväčší dôraz sme kládli na metriku RMSE, pretože nám spomedzi opísaných metrík najlepšie odráža chybu predikcie nehnuteľnosti. Sumár hodnotenia všetkých modelov ponúkame v nasledujúcich 2 tabuľkách:

| **Model** | **Train RMSE** | **Valid RMSE** | **Train RMSLE** | **Valid RMSLE** |
|---|---|---|---|---|
| Primitívny prediktor | 166,297 | 163,996 | 0.40446 | 0.44387 |
| Lin. regr. (sqft_living) | 137,783 | 142,066 | 0.33484 | 0.34669 |
| Lin. regr. (grade) | 135,747 | 140,737 | 0.3295 | 0.34308 |
| Lin. regr. (feat1)+filter | 129,314 | 135,827 | 0.31311 | 0.32878 |
| Lin. regr. (feat2) + SFS | 85,497 | 92,848 | 0.20066 | 0.22079 |
| Lin. regr. (feat3) + RFE | 83,407 | 91,041 | 0.19741 | 0.21855 |
| Poly. regr. (feat1) - st. 2 | 78,925 | 86,298 | 0.19039 | 0.20922 |
| Poly. regr. (feat1) - st. 3 | 52,234 | 87,008 | 0.13011 | 0.20563 |
| Regr. DT - max. depth 10 | 58,884 | 87,772 | 0.14407 | 0.21 |
| Regr. DT - max. depth 8 | 70,668 | 84,801 | 0.17105 | 0.21 |
| Random Forest trees=100 | 63,148 | 77,417 | 0.15331 | 0.19 |

**Table 3.** Výsledky použitých modelov pre metriky RMSE, RMSLE

Najlepšie výsledky podľa metriky RMSE dosiahol model Náhodného lesa ( 63,148 na trénovacej a 77,417 na validačnej množine), druhý najlepší výsledok podľa tejto metriky dosiahol model Regresného rozhodovacieho stromu (84,801 na trénovacej a 70,668 na validačnej množine).

V prípade modelu jednoduchej lineárnej regresie sme vyskúšali aj 5, resp. 10 - násobnú krížovú validáciu, na porovnanie výsledkov s modelom jednoduchej lineárnej regresie trénovaného tradičným spôsobom, avšak tento prístup nám nepreukázal žiaden signifikantný rozdiel vo výsledkoch pri modeloch jednoduchej lineárnej regresie.

| **Model** | **Train MSE** | **Valid MSE** | **Train $R^2$** | **Valid $R^2$** |
|---|---|---|---|---|
| Primitívny prediktor | 27,654,781,496 | 26,894,754,735 | - | - |
| Lin. regr. (sqft_living) | 18,984,293,278 | 20,182,781,602 | 0.311 | 0.258 |
| Lin. regr. (grade) | 18,427,517,960 | 19,807,129,202 | 0.332 | 0.273 |
| Lin. regr. (feat1)+filter | 16,722,154,445 | 18,449,215,523 | 0.396 | 0.328 |
| Lin. regr. (feat2) + SFS | 7,309,816,064 | 8,620,832,067 | 0.749 | 0.697 |
| Lin. regr. (feat3) + RFE | 6,956,876,768 | 8,288,558,829 | 0.759 | 0.706 |
| Poly. regr. (feat1) - st. 2 | 6,229,272,454 | 7,447,377,768 | 0.78 | 0.733 |
| Poly. regr. (feat1) - st. 3 | 2,728,439,519 | 7,570,530,881 | 0.901 | 0.76 |
| Regr. DT - max. depth 10 | 3,467,362,203 | 7,703,963,969 | 0.876 | 0.715 |
| Regr. DT - max. depth 8 | 4,994,066,550 | 7,191,050,598 | 0.823 | 0.74 |
| Random Forest trees=100 | 3,987,683,819 | 5,993,519,471 | 0.859 | 0.785 |

**Table 4.** Výsledky použitých modelov pre metriky RMSE, RMSLE a $R^2$

Model Náhodného lesa, ktorý sa v porovnaní všetkých modelov dosahoval najlepšie výsledky sme pomocou techník RandomSearch a následne GridSearch vyladili tak, aby sme získali čo najmenšiu hodnotu rozdielu chýb medzi Train RMSE a Valid RMSE a chyby RMSE pre validačnú množinu. Ako najoptimálnejšiu kombináciu hyperparametrov sme určili:

- maximálna hĺbka = 10
- maximálny počet atribútov (čŕt) = 18
- minimálny počet vzoriek pre listový vrchol = 1
- minimálny počet vzoriek pre rozdelenie vnútorného vrchola = 4
- počet stromov = 300

Model s takouto kombináciou hyperparametrov sme následne nechali predikovať vzorky na testovacích dátach a dosiahli sme takéto výsledky:

| **Metrika** | **Model** |
|---|---|
| Train MSE | 2,683,022,842.5495 |
| Valid MSE | 5,476,893,781.90784 |
| Test MSE | 6,227,529,395.53461 |
| Train RMSE | 51,797.90384 |
| Valid RMSE | 74,006.03882 |
| Test RMSE | 78,914.69696 |
| Train RMSLE | 0.12742 |
| Valid RMSLE | 0.18053 |
| Test RMSLE | 0.18116 |
| Train $R^2$ | 0.905 |
| Valid $R^2$ | 0.803 |
| Test $R^2$ | 0.769 |

**Table 5.** Výsledky odladeného modelu Náhodného lesa pre trénovaciu, validačnú a testovaciu množinu

Z výsledkov v 5 môžeme opäť pozorovať možné pretrénovanie, ktorému sme však ani minimalizáciou hodnôt hyperparametrov nedokázali v signifikantnej miere zabrániť. Pri porovnaní s modelmi vytvorenými v príbuzných prácach môžeme konštatovať, že náš model dosahoval podstatne lepšie výsledky ako v prípade [3], pri [1] sme bohužiaľ nedokázali určiť s akými jednotkami pri metrike MSE pracovali autori.

## 6 Zhrnutie

V tomto projekte sme sa zamerali na regresné modely, pretože naším cieľom bolo vytvoriť kvalitný model predikcie cien nehnuteľností. Tento cieľ sa nám podarilo splniť, po vykonaní experimentov sa ukázalo, že najlepší model pre náš prípad bude náhodný les. Počas práce na projekte sme sa stretli s viacerými nástrahami dátovej analýzy, ako napríklad čistenie dát, feature engineering, normalizácia, ladenie hyperparametrov a iné. Vyskúšali sme viacero modelov a všetky dosiahnuté výsledky zdokumentovali.

Ďalšie smerovanie tohto projektu vidíme v možnej optimalizácii predspracovania dát, konkrétne experimentovanie s transformáciou odhadovaného atribútu price, poprípade experimentovaní s inými, viac exotickými regresnými modelmi.

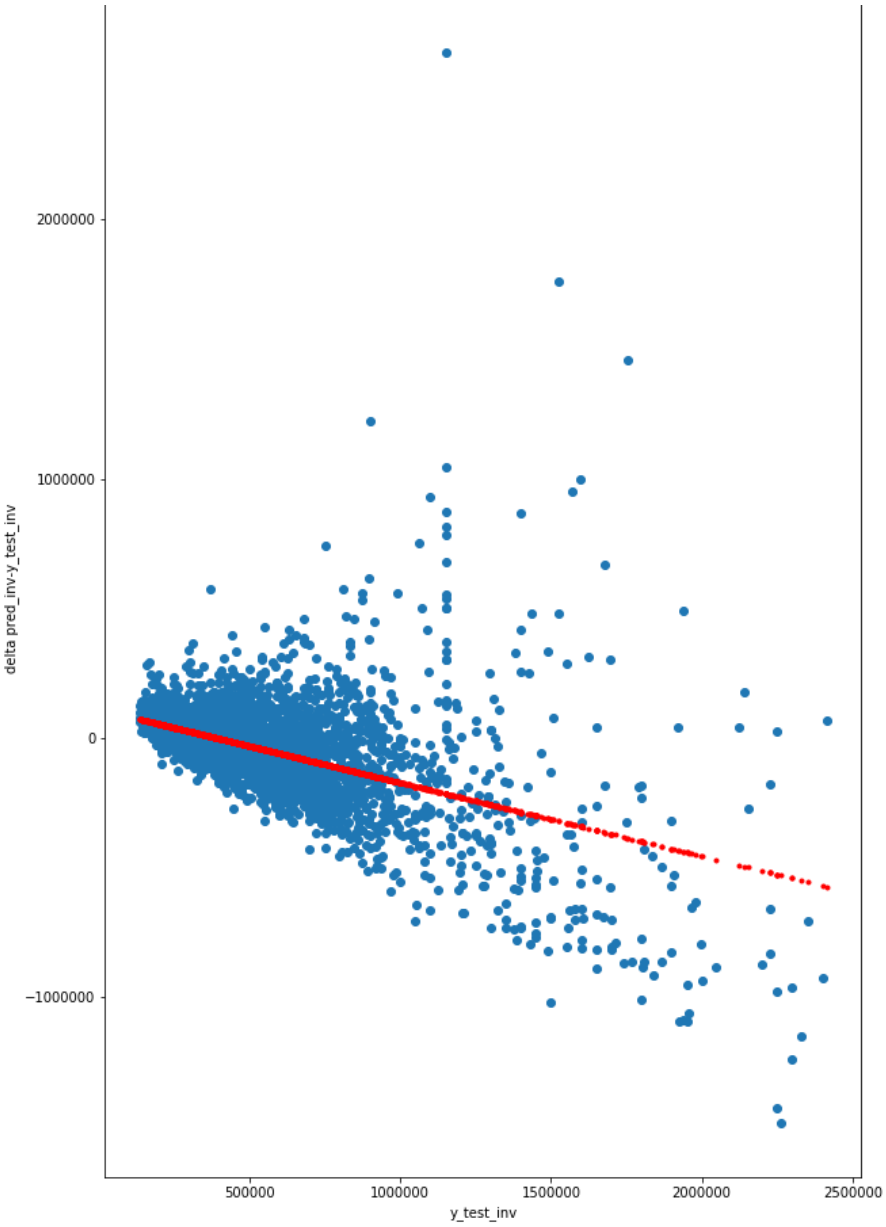


**Fig. 1.** Grafické zobrazenie závislosti skutočnej hodnoty a rozdielu predikovanej a skutočnej hodnoty ceny. Z grafu si môžeme všimnúť rastúci rozptyl rozdielov predikovaných cien so zvyšujúcou sa skutočnou cenou nehnuteľností

# References


[1] Phan, The Danh: Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia. In: 2018 International Conference on Machine Learning and Data Engineering (iCMLDE), pp. 35–42., IEEE (2018)

[2] Xue, Chao and Ju, Yongfeng and Li, Shuguang and Zhou, Qilong: Research on the Sustainable Development of Urban Housing Price Based on Transport Accessibility: A Case Study of Xi'an, China. Sustainability **12**(4), 1497–1513 (2020)

[3] Mutyala, Nikhil Kumar, `https://towardsdatascience.com/regression-using-sklearn-on-kc-housing-dataset-1ac80ca3d6d4`. Last accessed 17 Apr 2020