TODO

įdėti nuotrauką su grotomis ir “grafiti” - kaip baisiai atrodo

SALE OR RENT PRICE PREDICTION OF HOUSES IN MADRID

SDA - DataScienceRemoteLT1 (Kęstutis, Alina, Lina)

Įrankiai

1. **G Colab** (3 instances of the same colab, and 2 …)
2. **anydesk**:
    a. **screen sharing**
    b. **clipboard sharing**
    c. **3 mouse pointers sharing**
3. **messenger** (**better sound** than zoom)
4. **Google Docs** (for **slides**)
5. mandatory **Zoom** (**screen sharing** and **talk with teacher**)

Spain, Madrid real estate data

8 MB: 21742 rows * 58 columns

Population:

1. Capital city and Municipality **3_2**23_334
2. Density 5_300/km^2
3. Urban **6_3**45_000
4. Metro **6_7**91_667

Euros from 1999.

Data updates **2 years ago**. Last updated 2020-04-18 and today is 2022-02-13.

Date Coverage. Temporal coverage 202**0-03-10** to 202**0-03-28 (18 dienų)**

Aprašėme 56 feature’us

1. NaN reikšmių kiek
2. pavyzdinius duomenis
3. apsirašėme kiek kokių reikšmių yra **TIK JEI** mažas unikalių reikšmių kiekiui
4. min-max
5. išsivertėme iš ispanų kalbos pagrindines reikšmes

“Real estate listings in Madrid crawled from popular internet portal**s**”

Bet **svarbus** feature’is “portal” neturi reikšmių

t. y. 100 %-21742 NaN (Not a Number) reikšmės

Išsirinkome 6 feature’us pagal koreliaciją iš 15 int feature’ų

int'ų tipo feature'ų yra 27 % (<-15/56)

SWOT (Project – „Sale or rent Price prediction of houses/flats in Madrid“)

SILPNYBĖS:

- Komanda yra suformuota nevienodo lygio specialistų – projektas nebus įgyvendintas;
- Tokio pobūdžio NT duomenys jau buvo nagrinėti paviršutiniškai - neįdomu;
- Šiuo metu Madrid‘o kainų prediction‘o modelio rezultatais nepasinaudosime.
- Temos problematika – tema nėra aktuali savo tiesioginiu poveikiu mums tol, kol neiškils tokio pobūdžio klausimas;
- Madrido miesto (Ispanija) gyventojų poreikių ir prioritetų pasirinkimas NT rinkoje nežinomas – lietuviui neaišku - analizė apsunkinta.

STIPRYBĖS:

- Komandos nariai susipažins artimiau, išmoks dirbti komandoje geriau (procesai, įrankiai), sukurs plačiau panaudojamą modelį/karkasą, patirtimi pasidalins su SDA kolegomis;
- Sukurtas modelis prognozuos (gal ir nuomos) kainą namo/buto, bus išbandytas jo tikslumas (patikimumas);
- Projekto tema ir parametrai nesudėtingi (neįtraukiant ispanų kalbos) suprasti visiems – nereikia papildomų žinių;
- Greičiausiai niekas nėra daręs tokio projekto pagal sudėties kombinaciją: programų įrankiai, komandos narių skirtingos patirtys ir kompetencija - autentiškumas.

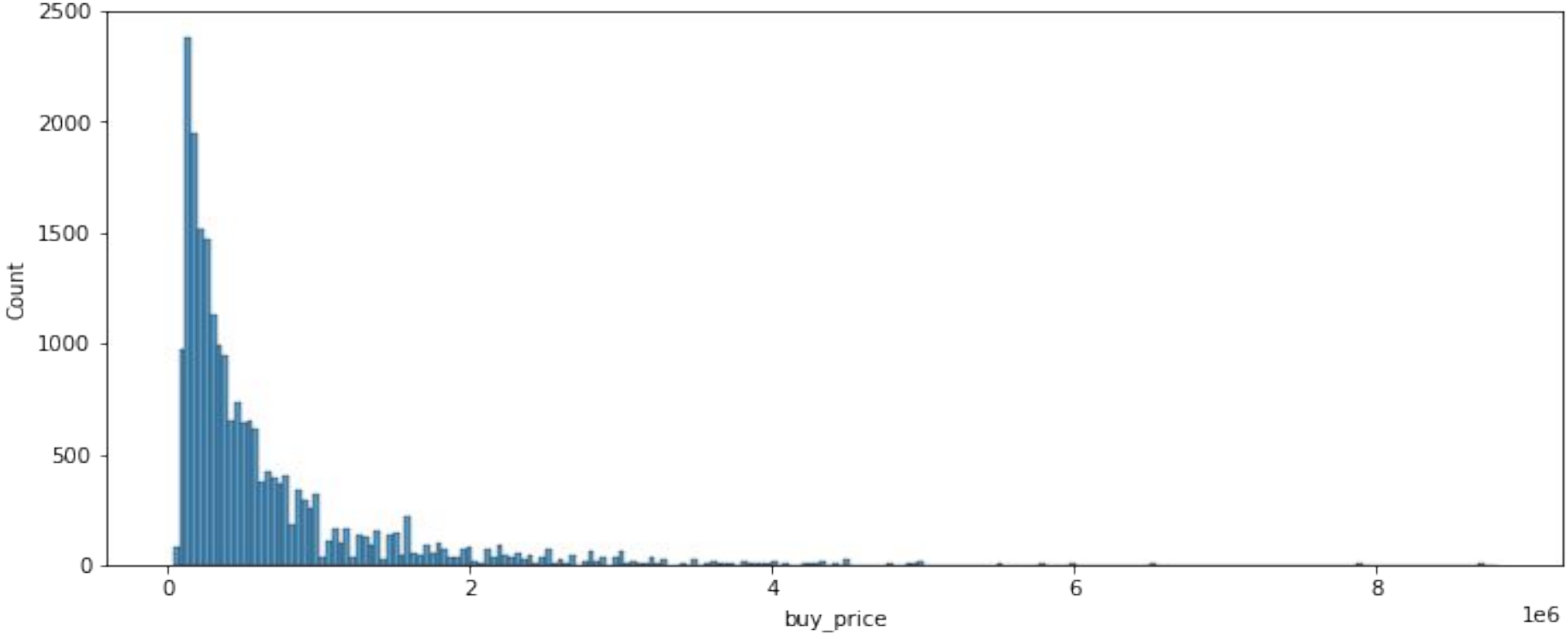
GRĖSMĖS:

- Komandos negebėjimas dirbti drauge ar pasikliauti vienas kitu neatneš norimo rezultato;
- Techninės kliūtys ir laiko trūkumas – nebaigtas ar nekokybiškas modelis;
- Per daug netinkamų (NaN, outlier(s)) – duomenys iškreips prediction‘o rezultatus;
- Netinkamas duomenų supratimas, interpretavimas duos klaidinančią išvadą;
- Sukurtas modelis neveiks.

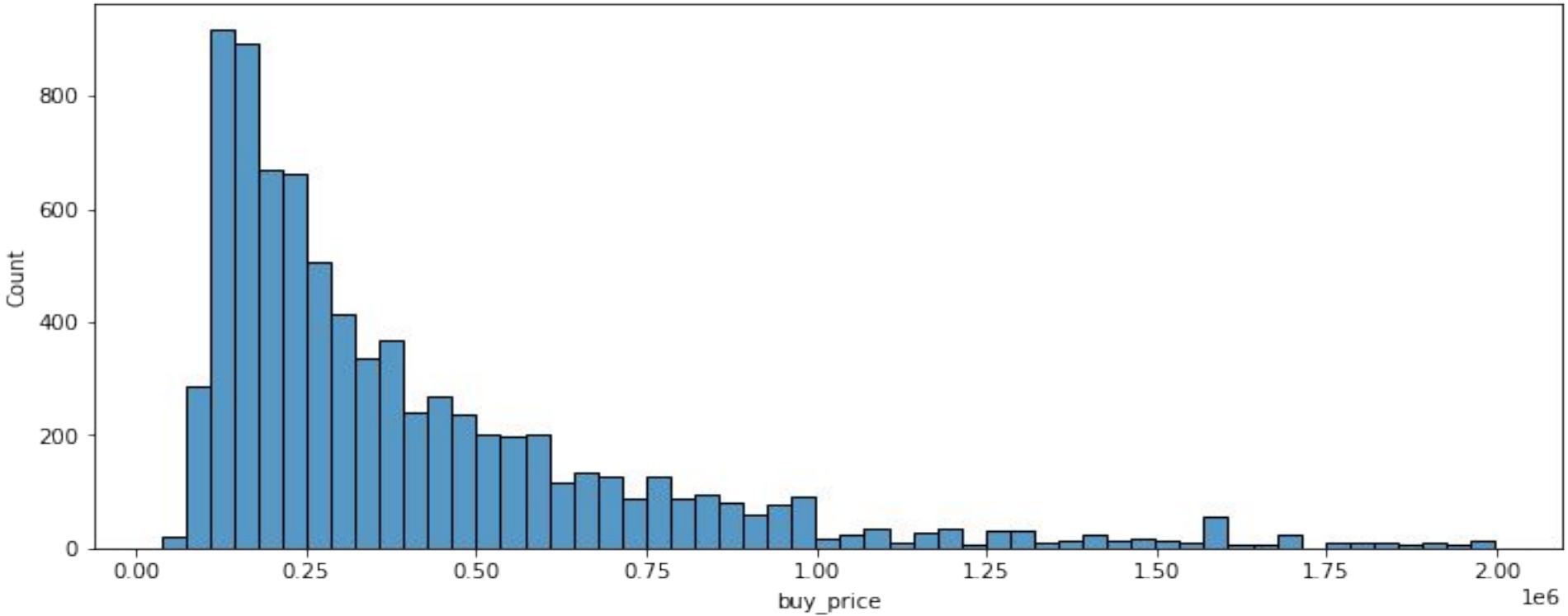
GALIMYBĖS:

- Komanda išmoks geriau: planuoti laiką, pasidalinti darbais, pažins vienas kitą, įgis praktinės patirties tokio tipo projektuose, gebės gautus rezultatus teisingai apibendrinti ir perteikti jų aktualumą SDA kolegoms;
- Projekto rezultatai gali nustebinti savo išvada – netikėtumas atradus naujas sąsajas NT feature‘ų ir „price“ atžvilgiu;
- Galimybės panaudoti modelį nagrinėjant kitų miestų NT duomenis – platus panaudojamumas.

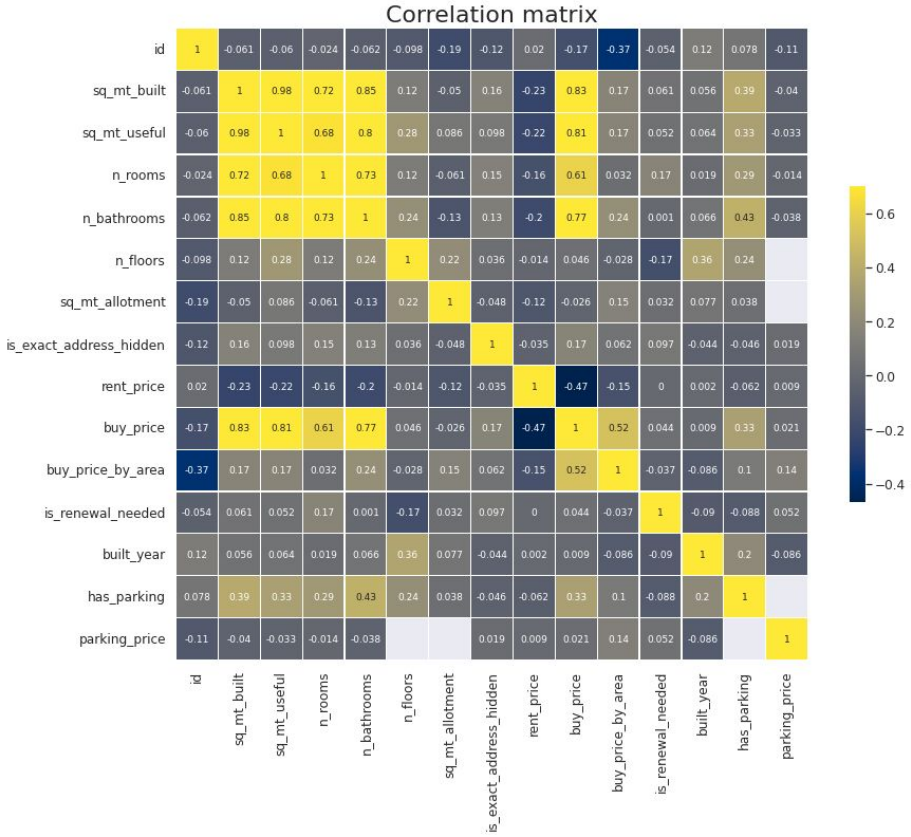
Visų būstų kainos histograma



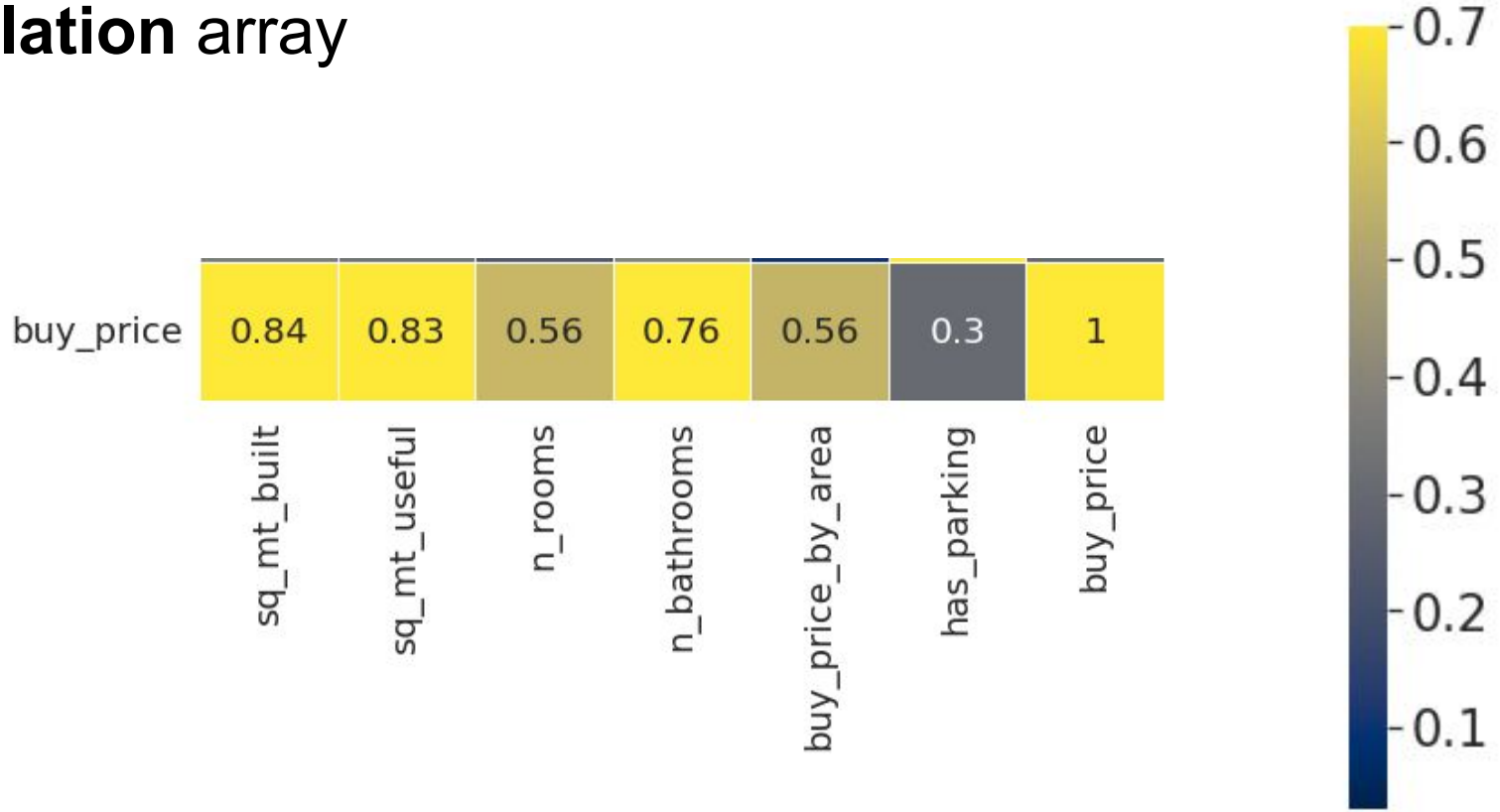
Apribota būsto kainos histograma

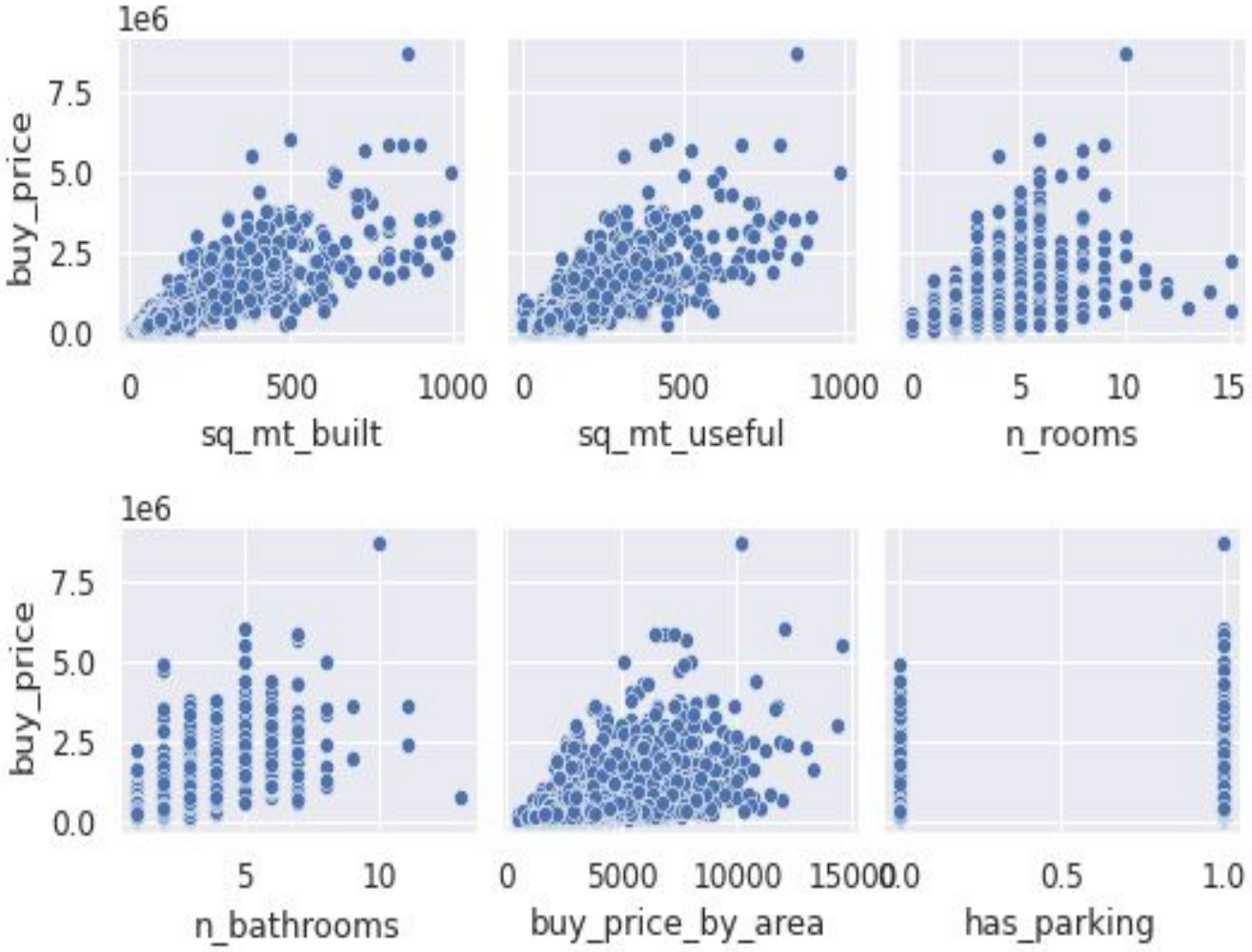


15 int tipo feature’ų koreliacija



Correlation array




1e6
buy_price
7.5
5.0
2.5
0.0
0
500
1000
sq_mt_built
sq_mt_useful
0
5
10
15
n_rooms
n_bathrooms
0
5000
10000
15000
buy_price_by_area
0.0
0.5
1.0
has_parking

Linear Regression

- **Buy price** < 2_000_000 eur

- **Features:** 1. sq_mt_built, 2. sq_mt_useful, 3. n_rooms, 4. n_bathrooms, 5. buy_price_by_area, 6. has_parking

- **Data split**: 70% (train), 30 % (test)

MAPE, WMAPE, RMSE

MAPE: 0.1948

WMAPE: 0.1474

RMSE: 105358.9387

Išvados

1. Nemokant ispanų kalbos, turimų duomenų interpretavimas gali būti klaidingas, nes gali neužtekti google translate vertimo ir neturime ispanakalbio.
2. Pasirinkti duomenys turi daug NaN, apsunkina analizę ir apdorojimą.
3. Laiko planavimas:
    a. Reikia žadintuvų kas tam tikrą laiką ir kas tam tikrą etapą.
    b. Planuoti laiko pasiskirstymą kiek kam užtrunkam.
    c. Fiksuoti įvykius (užduočių pradžią ir pabaigą)
4. Geriau pasiskirstyti darbus, labiau sudėlioti atsakomybes.
5. Būtina didesnis NT rinkos analizė (tiek feature'ų tiek rinkos specifikos iš kitų analizių ir iš pačių rinkos dalyvių kalbant su jais) kad galėtume teisingai įvertinti gautus duomenis ir padaryti tikslias išvadas ir jas patikrinti.
6. Galima toliau tikslinti tikslumą:
    a. padidinant feature'ų kiekį
    b. naudojant geresnį algoritmą