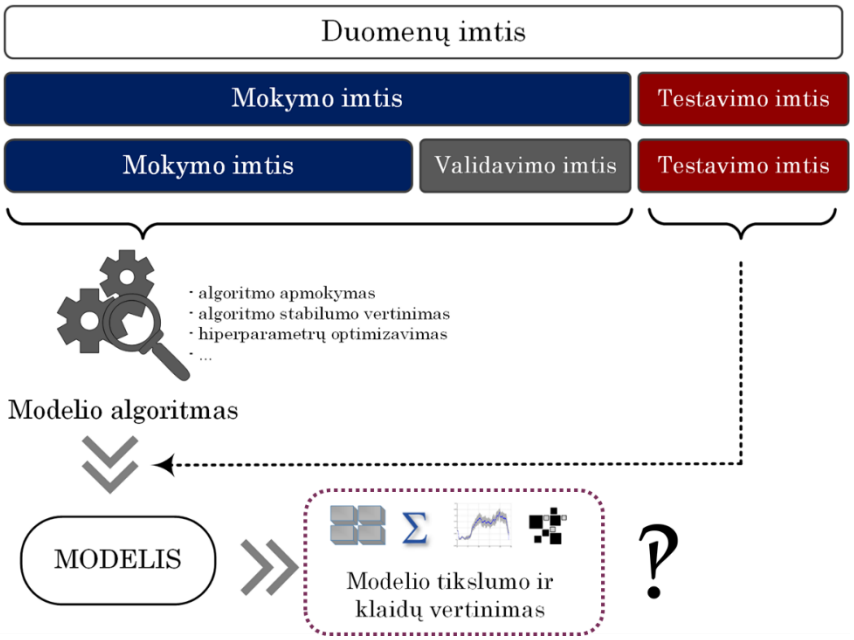
Prižiūrimojo mašininio mokymosi modelio patikra

III dalis. Prižiūrimojo mašininio mokymosi modelio patikra, kai tikslinis kintamasis yra tolydusis (regresijos modelis)

Dar kartą prisiminkime jau matytą modelio patikros principinę schemą (1 pav.), kuri buvo pristatyta pirmoje šios temos dalyje.



1 pav. *Modelio patikros schema*

Šioje temoje pagrindinį dėmesį skiriame paskutiniam etapui, kai norime įvertinti modelio prognozavimo tikslumą ir išsirinkti geresnį modelį iš kelių galimų.

Atliekant modelio patikrą, kai tikslinis kintamasis yra tolydusis, pagrindinis dėmesys skiriamas ***klaidos kintamajam*** $e_i,\ i = \overline{1,m}$, kuris šiuo atveju kur kas dažniau vadinamas tiesiog ***paklaida***. Kintamasis e_i apskaičiuojamas validavimo / testavimo imtims kaip skirtumas tarp tikslinio kintamojo faktinių reikšmių y_i ir modelio gautų prognozių $\hat{y}_i$ pagal formulę

$$e_i = y_i - \hat{y}_i,\ \ i = \overline{1,m}.$$

Rekomenduojama rinktis tą modelį, kurio paklaidos yra mažesnės. Tai kyla klausimas, kaip įvertinti paklaidų e_i didumą ar mažumą?

Paklaidų dydį vertiname skaičiuodami įvairias paklaidų metrikas. Gaunamas vienas skaičius (įvertis), pagal kurį sprendžiame apie modelio tinkamumą ar renkamės geresnį modelį iš kelių galimų.

Pirma metrikų grupė yra ***masteliui jautrios paklaidų metrikos***. Tarp populiariausių yra:

- ***vidutinė absoliutinė paklaida*** (angl. *Mean absolute error*, MAE)

$$\text{MAE} = \frac{1}{m}\sum_{i=1}^{m}|e_i|$$

- ***vidutinė kvadratinė paklaida*** (angl. *Mean squared error*, MSE)

$$\text{MSE} = \frac{1}{m}\sum_{i=1}^{m} e_i^2$$

- ***šaknis iš vidutinės kvadratinės paklaidos*** (angl. *Root mean squared error*, RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m} e_i^2}$$

Šių paklaidų didumas gan stipriai priklauso nuo to, koks yra tikslinio kintamojo mastelis, t. y. šimtai, tūkstančiai ar milijonai.

Kita metrikų grupė yra ***procentinės paklaidų metrikos***. Tarp populiariausių yra:

- ***vidutinė absoliutinė procentinė paklaida*** (angl. *Mean absolute percentage error*, MAPE)

$$\text{MAPE} = \frac{1}{m}\sum_{i=1}^{m}\left|\frac{e_i}{y_i}\right| \cdot 100\%$$

- ***simetrinė absoliutinė procentinė paklaida*** (angl. *Symmetric mean absolute percentage error*, sMAPE)

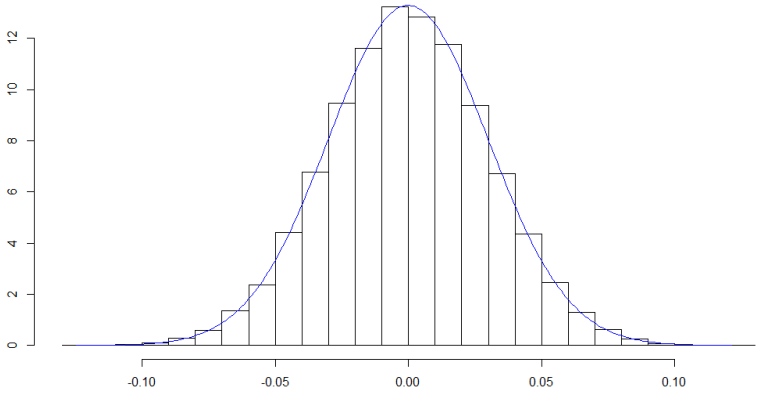
$$\text{sMAPE} = \frac{1}{m}\sum_{i=1}^{m}\left|\frac{e_i}{y_i + \hat{y}_i}\right| \cdot 100\%$$

Gan dažnai modelio gerumui vertinti tyrėjai renkasi ir ***determinacijos koeficientą*** R^2. Jis tarsi paaiškina, kaip gerai sudarytas modelis aprašo ar aproksimuoja duomenis. Determinacijos koeficientas apskaičiuojamas pagal formulę

$$\text{R}^2 = \frac{\sum_{i=1}^{m}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{m}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{m}(y_i - \bar{y})^2}$$

Determinacijos koeficiento įgyjamos reikšmės yra iš intervalo $(0; 1)$. Kuo reikšmė artimesnė 1, tuo modelis tinkamesnis sprendžiamam uždaviniui.

Papildomai verta atkreipti dėmesį į gautų modelio paklaidų e_i, $i = \overline{1, m}$ histogramą, kuri savo forma turėtų būti panaši į normalųjį (Gauso) pasiskirstymą (2 pav.).



2 pav. Normalusis (Gauso) skirstinys

Panagrinėkime pavyzdį. Tarkime, kad modeliui testuoti atidėta 20 stebėjimų. Lentelėje pateiktas jo numeris ir faktinė reikšmė (3 pav.). Sakykime, kad buvo išmokyti du modeliai A ir B, o jų gautos prognozės pateiktos (3) ir (5) stulpeliuose. Atitinkamai (4) ir (6) stulpeliuose yra pateiktos apskaičiuotos paklaidos, kaip skirtumas nuo faktinės ir prognozuotos reikšmių.

ID	Faktas, y_i	„A“ modelio prognozė, $\hat{y}_i$	„A“ modelio paklaida, e_i	„B“ modelio prognozė, $\hat{y}_i$	„B“ modelio paklaida, e_i
(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
M1	2,7	2,5	0,2	2,4	0,3
M2	2,6	2,4	0,2	2,1	0,5
M3	1,7	1,9	-0,2	1,9	-0,2
M4	1,8	1,6	0,2	1,5	0,3
M5	2,2	2,0	0,2	2,0	0,2
M6	2,8	2,7	0,1	2,4	0,4
M7	3,1	2,9	0,2	2,7	0,4
M8	3,2	3,0	0,2	2,8	0,4
M9	3,8	3,5	0,3	2,9	0,7
M10	2,7	2,8	-0,1	2,8	-0,1
M11	2,0	2,0	0,0	2,4	-0,4
M12	1,2	1,5	-0,3	2,2	-0,7
M13	2,5	2,3	0,2	2,3	0,2
M14	2,7	2,6	0,1	2,5	0,2
M15	1,8	2,2	-0,4	2,4	-0,6
M16	2,7	2,8	-0,1	2,9	-0,2
M17	2,5	2,7	-0,2	3,1	-0,6
M18	2,8	2,9	-0,1	2,6	0,2
M19	3,2	3,0	0,2	2,9	0,3
M20	2,6	2,7	-0,1	2,7	-0,1

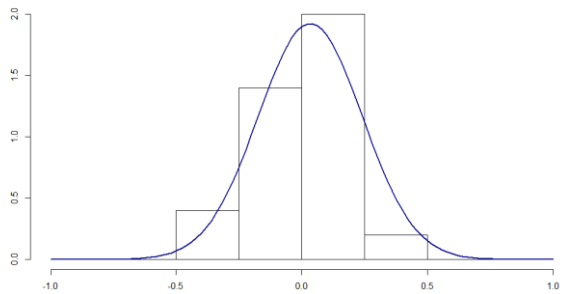
3 pav. *Demonstracinis pavyzdys: A ir B modelių prognozės ir apskaičiuotos paklaidos kiekvienam stebėjimui*

Paklaidų metrikos apskaičiuotos pagal pateiktas formules ir gauti jų įverčiai surašyti lentelėje. Pastebėsime tai, kad A modelio visos paklaidų metrikos yra mažesnės lyginant su B modeliu, o determinacijos koeficientas yra didesnis.

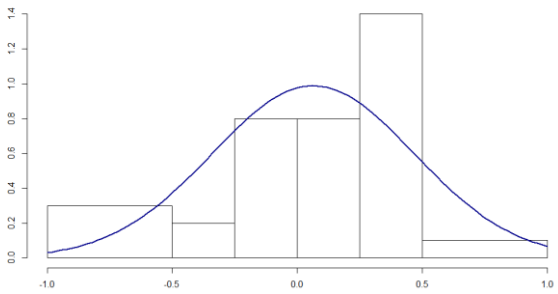
	„A“ modelis	„B“ modelis
Apskaičiuotos paklaidų metrikos		
ME	0,03	0,06
MAE	0,18	0,35
MSE	0,04	0,16
RMSE	0,20	0,39
MAPE, %	7,96	15,43
sMAPE, %	7,80	14,79
Apskaičiuotas determinacijos koeficientas		
R^2	0,71	0,51

4 pav. *Paklaidos ir determinacijos koeficientas*

Šiuo atveju „A“ modelis tinkamesnis nei „B“ modelis. Atvaizduokime abiejų modelių paklaidas histograma.



5 pav. *A modelio paklaidų histogram ir Gauso kreivė*



6 pav. *B modelio paklaidų histograma ir Gauso kreivė*

Trumpas komentaras dėl hiperparametrų parinkimo. Tolydžiojo tikslinio kintamojo reikšmėms prognozuoti gali būti taikomi tokie regresiniai mašininio mokymo metodai, kurie turi hiperparametrus. Šių parametrų optimalios reikšmės randamos vertinant paklaidas validavimo imčiai. Pavyzdžiui, regresiniame atsitiktinių miškų metode hiperparametras yra medžių skaičius miške. Kitas pavyzdys, istorinių stebėjimų, naudojimų modelio apmokymui, kiekio kalibravimas.

Apibendrinant:

- geresnis modelis yra tas, kuris turi mažesnes paklaidas validavimo / testavimo imčiai;
- šioje temoje aptartos dažniausiai naudojamos paklaidų metrikos, t. y. MAE, MSE, RMSE, MAPE, sMAPE. Paklaidas siekiama minimizuoti ieškant geriausio modelio. Literatūros šalininiuose siūloma ir daugiau įvairesnių paklaidų metrikų;
- jei taikomas determinacijos koeficientas R^2, tai geresnis modelis yra tas, kurio R^2 yra didesnis.