

ерекшелігі инженериясы
және үлгіні таңдау

ML рецепті

1. Кейбір үлгіні ойлап көріңіз
2. Деректерді үлгіге енгізіп, болжам жасаңыз.
3. Болжамдар мен шынайы мәндер арасындағы шығынды есептеңіз.
4. Ең аз шығын келтіретін үлгі параметрлерін анықтаңыз.

ML рецепті

1. Кейбір үлгіні ойлап көріңіз
2. Деректерді үлгіге енгізіп, болжам жасаңыз.
 1. Біз қандай деректерді қосу керектігін шешеміз.
 2. Біз деректерді мүмкіндіктерге қалай айналдыру керектігін шешеміз.
3. Болжамдар мен шынайы мәндер арасындағы шығынды есептеңіз.
4. Ең аз шығын келтіретін үлгі параметрлерін анықтаңыз.

Инженерлік мүмкіндіктер:

Деректерді мүмкіндіктерге
айналдыру

Сызықтық модельдер үшін сызықты емес мүмкіндіктер

- Біз X мүмкіндіктеріне қалағанымызды жасай аламыз.

$$y_i = \sum_{j=0}^p \beta_j X_{ij}$$

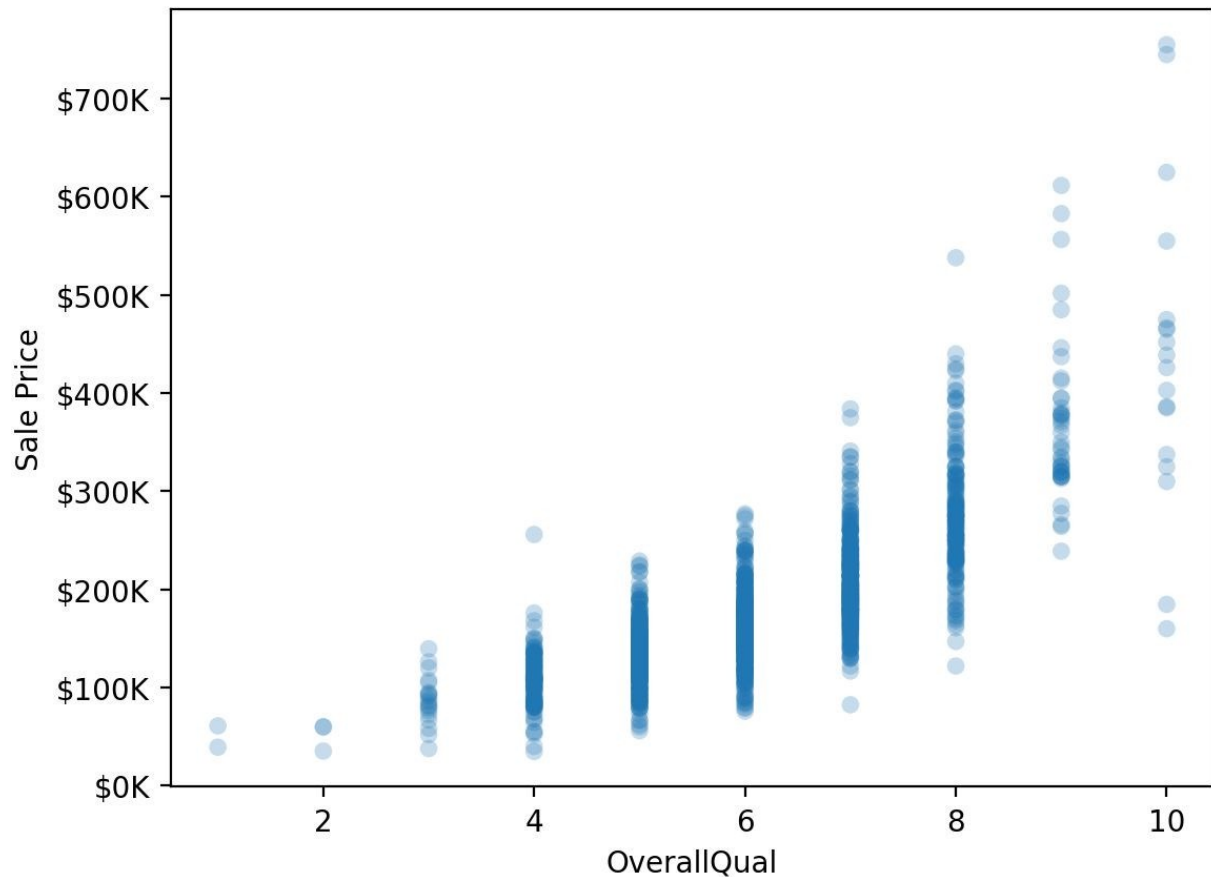
- Біз бір функцияны квадраттай аламыз, біз бір-бірімізбен бірнеше функцияларды жасай аламыз, синусты қолдана аламыз және т.б.

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i1}^2 + \beta_3 X_{i1} X_{i2} + \beta_4 \sin(X_{i3})$$

- Бұл мүмкіндіктер сызықты емес болғанымен, модель параметрлері β бойынша сызықты болады.

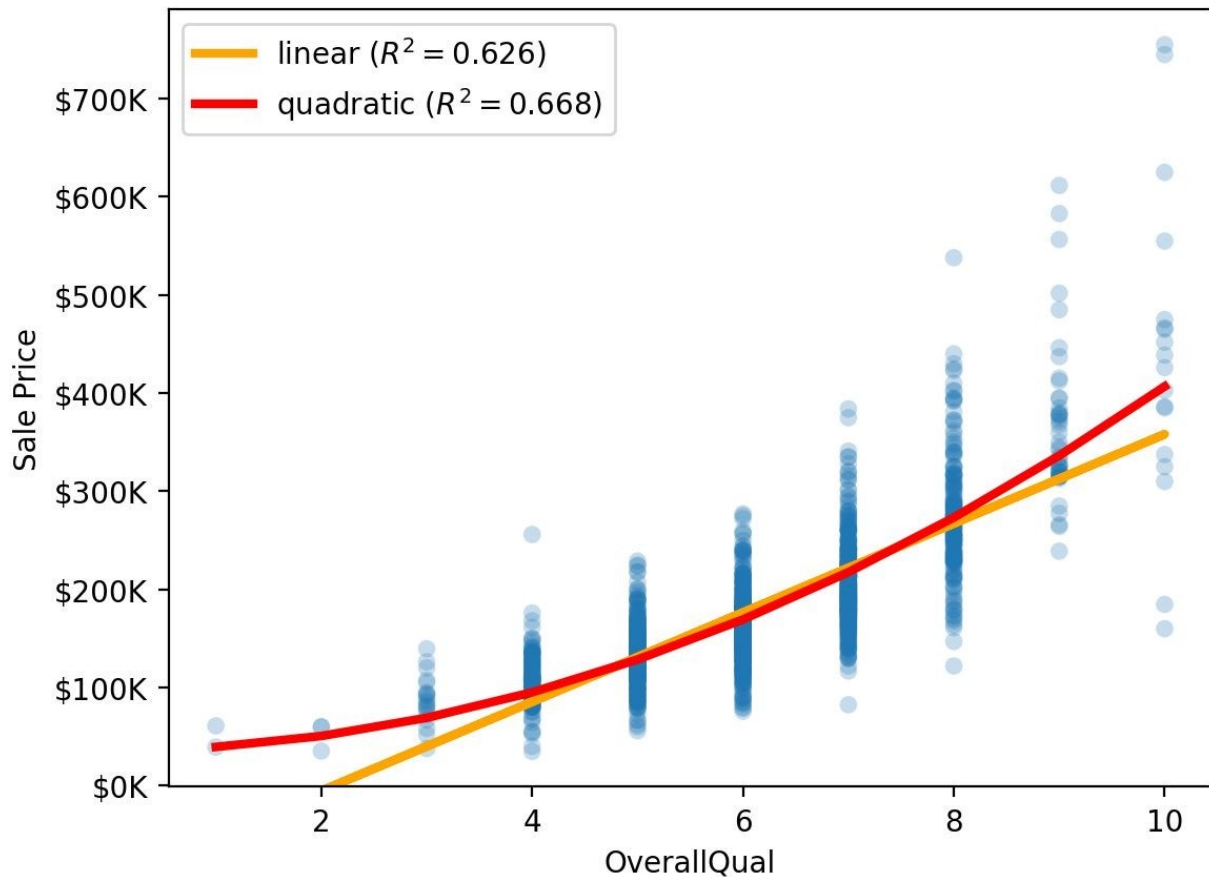
Неліктен инженерлік мүмкіндіктер?

Біз жақсырақ білеміз
деректер <> мақсатты
қатынас



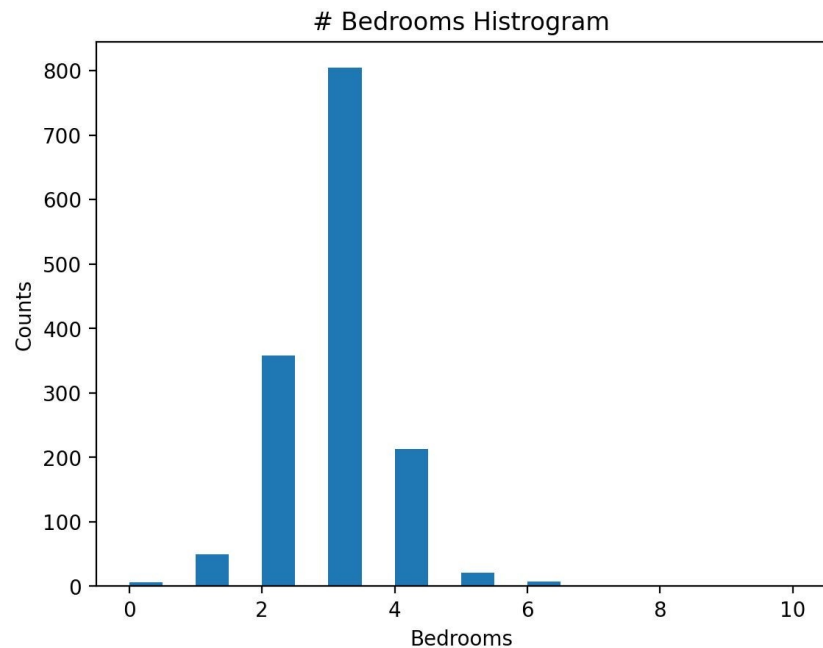
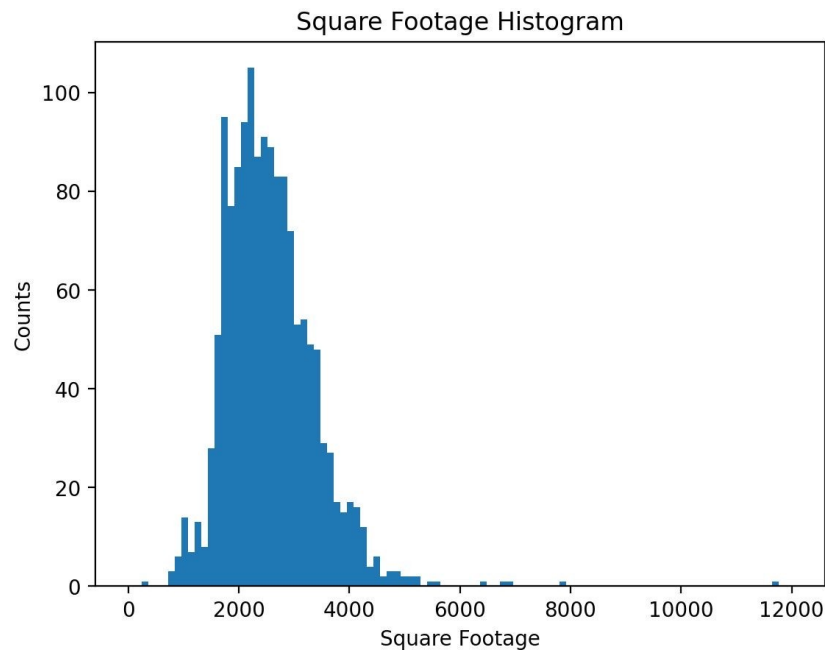
Неліктен инженерлік мүмкіндіктер?

Біз жақсырақ білеміз
деректер <> мақсатты
қатынас



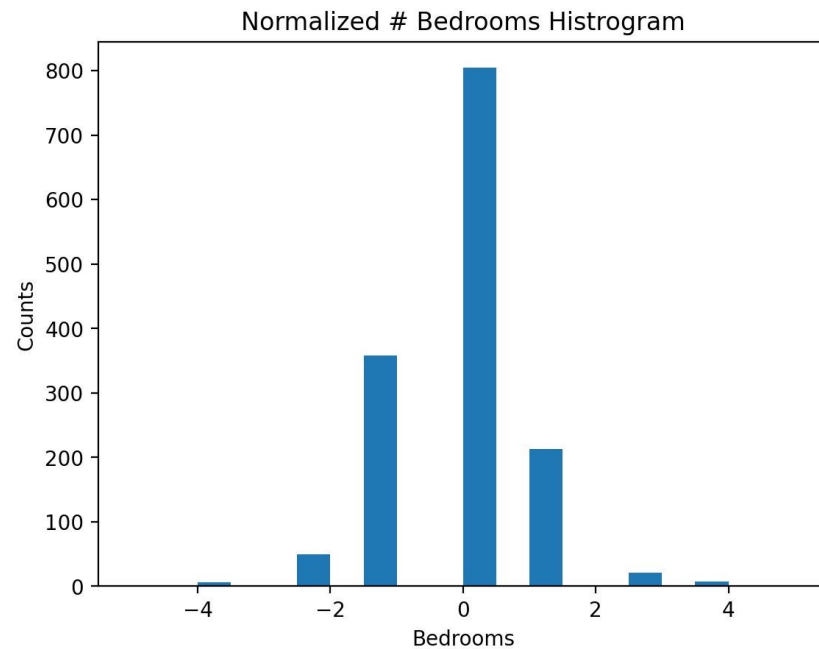
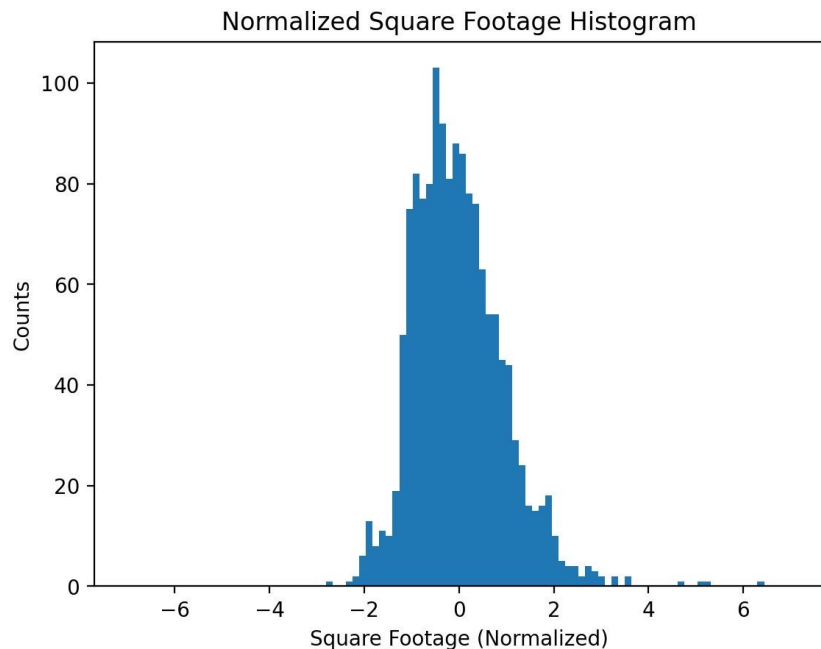
Неліктен инженерлік мүмкіндіктер?

Мүмкіндіктерді бірдей шкалаға қойыңыз



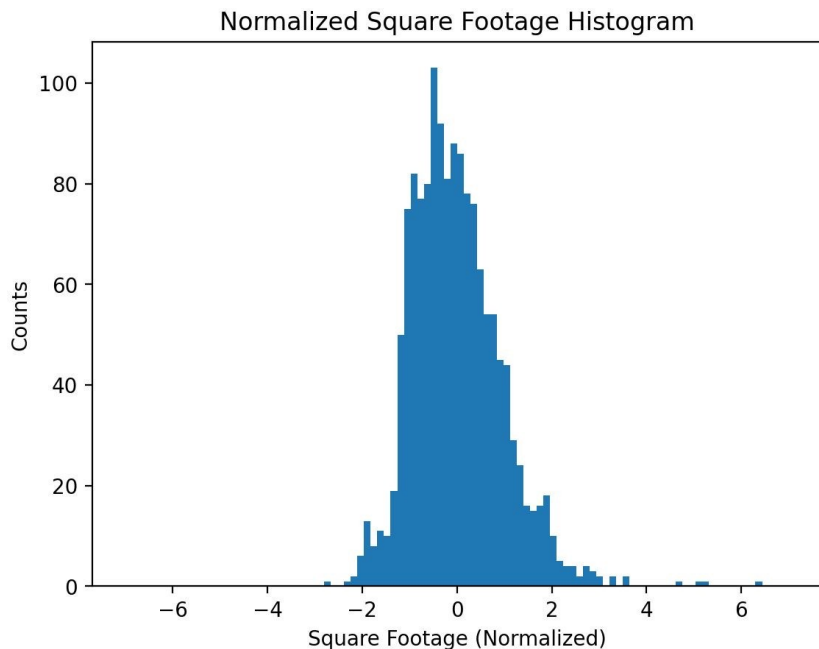
Неліктен инженерлік мүмкіндіктер?

Мүмкіндіктерді бірдей шкалаға қойыңыз



Неліктен инженерлік мүмкіндіктер?

Мүмкіндіктерді бірдей шкалаға қойыңыз



Normalization: $\vec{\mathbf{X}}_j^* = \frac{\vec{\mathbf{X}}_j - \bar{X}_j}{VAR(\vec{\mathbf{X}}_j)}$

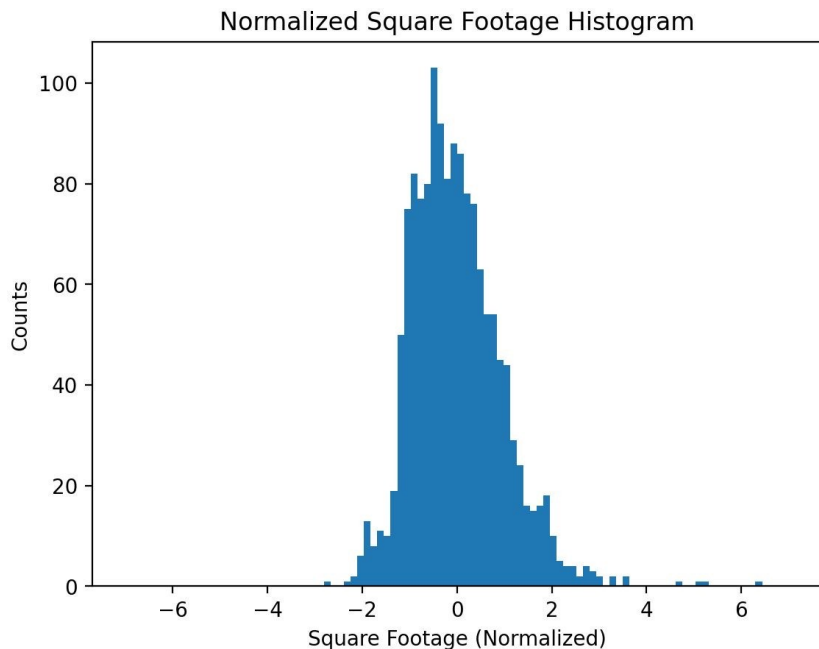
where

$$\bar{X}_j = \frac{1}{n} \sum_{i=1}^n X_{ij}$$

$$VAR(X_j) = \frac{1}{n-1} \sum_{i=1}^n (X_{ij} - \bar{X}_j)^2$$

Неліктен инженерлік мүмкіндіктер?

Мүмкіндіктерді бірдей шкалаға қойыңыз

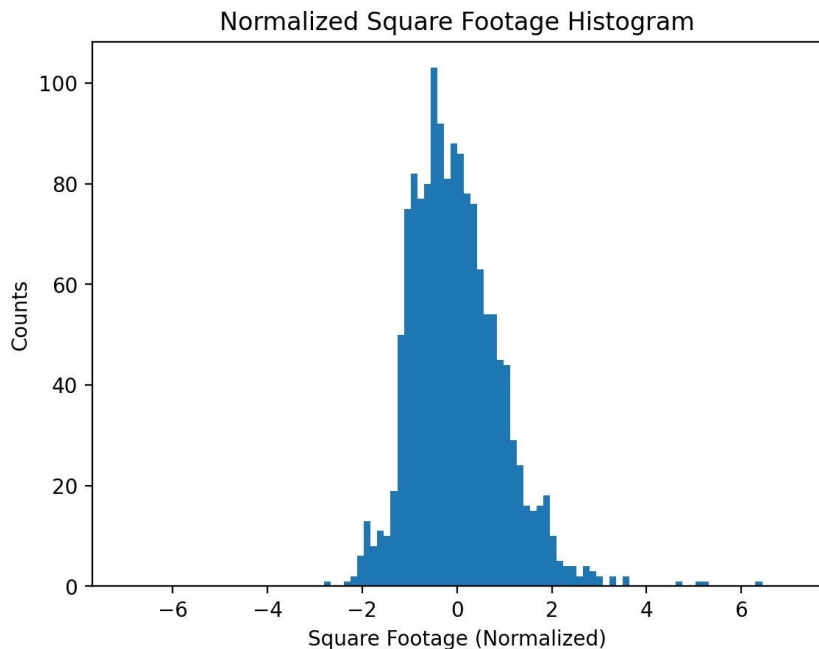


Көптеген басқа мүмкіндіктерді
масштабтау әдістері:

- Лог-преобразование
- Мин/Макс
масштабирование
- Макс Абс
масштабирование
- Степень преобразования

Неліктен инженерлік мүмкіндіктер?

Мүмкіндіктерді бірдей шкалаға қойыңыз

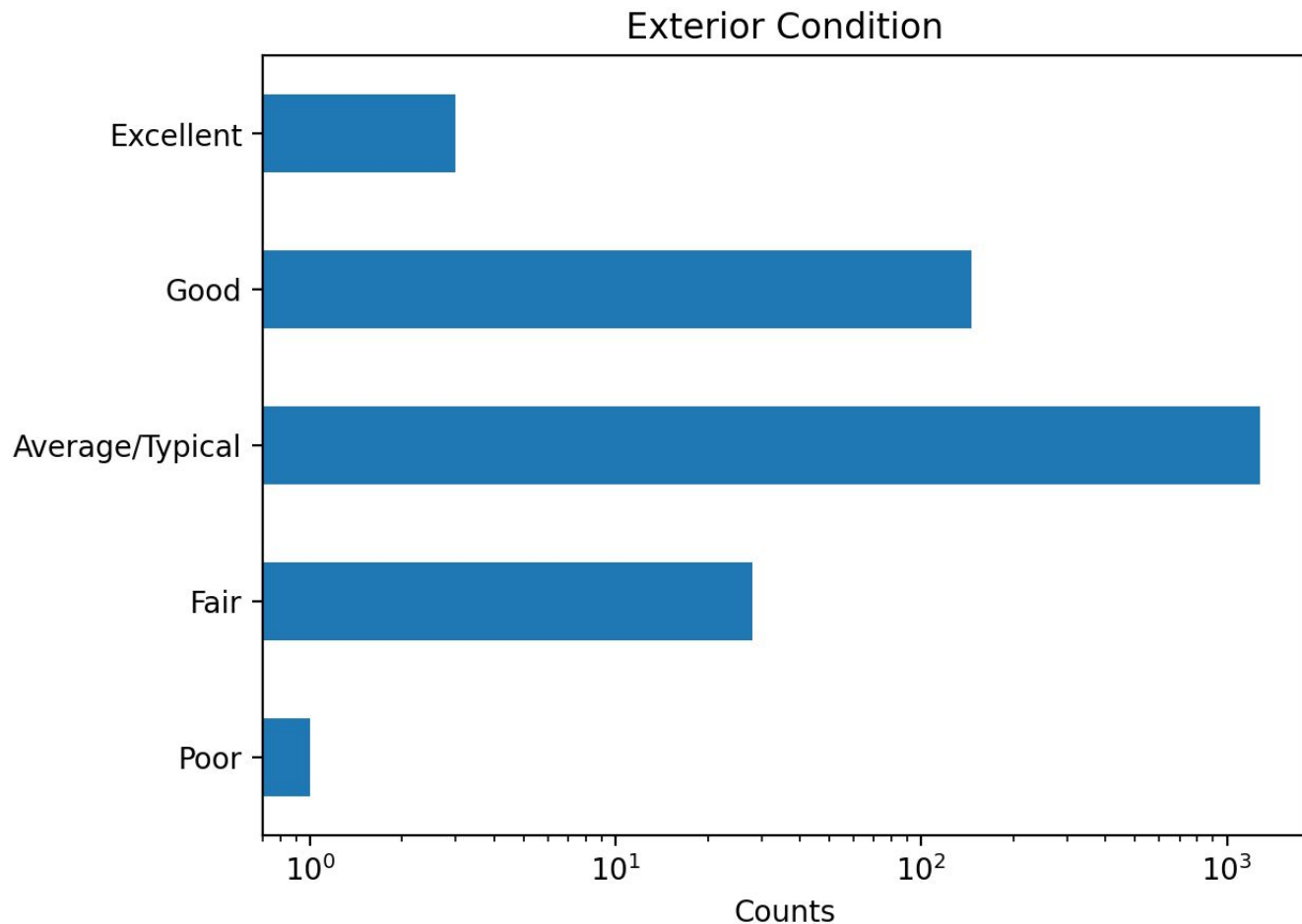


Неліктен мүмкіндіктерді бір шкалаға қою керек?

- Сызықтық модельдерден қорытынды жасай алады.
- Кейбір алгоритмдер тезірек біріктіріледі.
- Кейбір алгоритмдер мүмкіндіктер масштабталған жағдайда ғана біріктіріледі.

Неліктен инженерлік мүмкіндіктер?

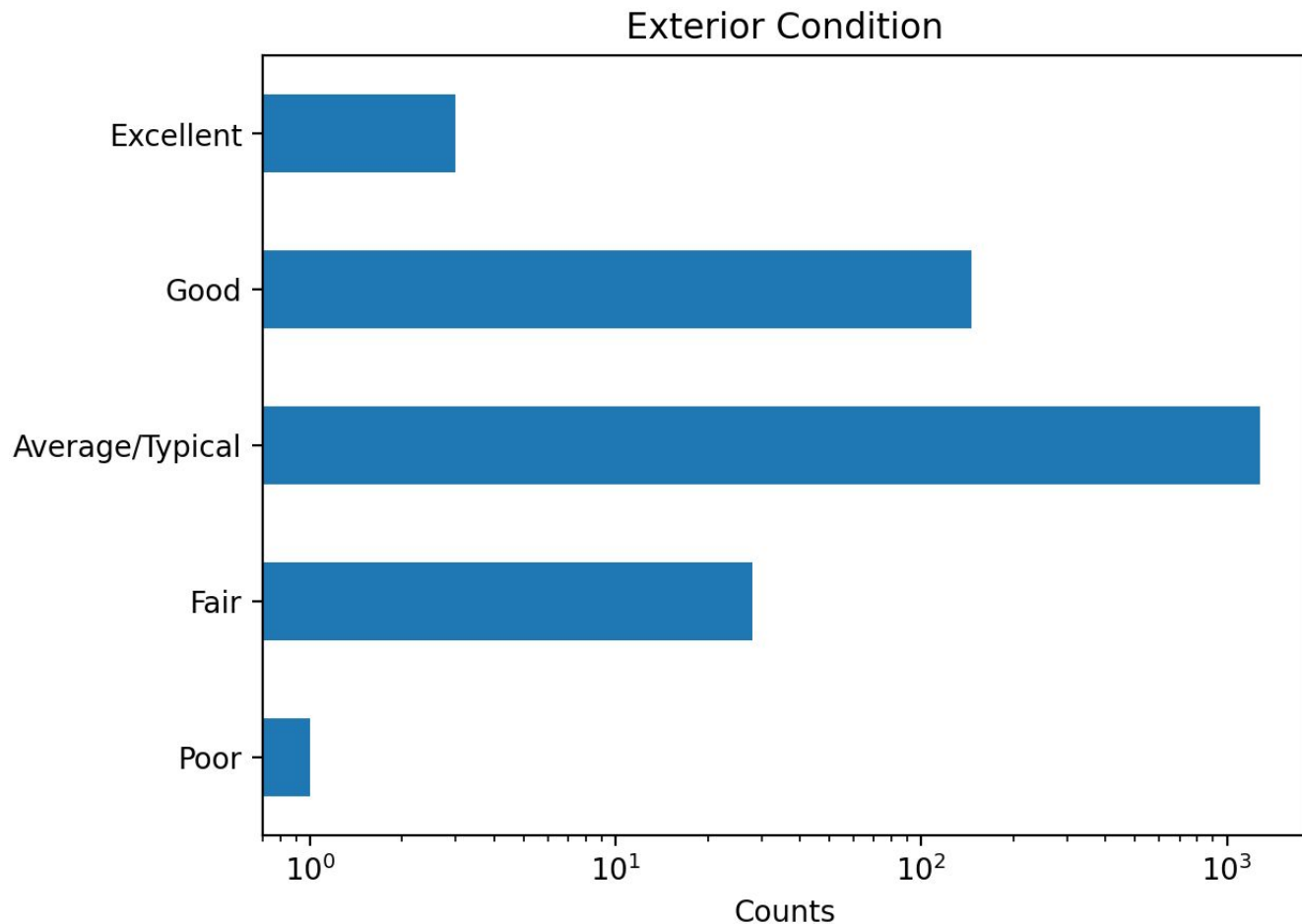
Кейде бізге қажет:



Барлық мүмкіндіктер
сандар болуы керек

Неліктен инженерлік мүмкіндіктер?

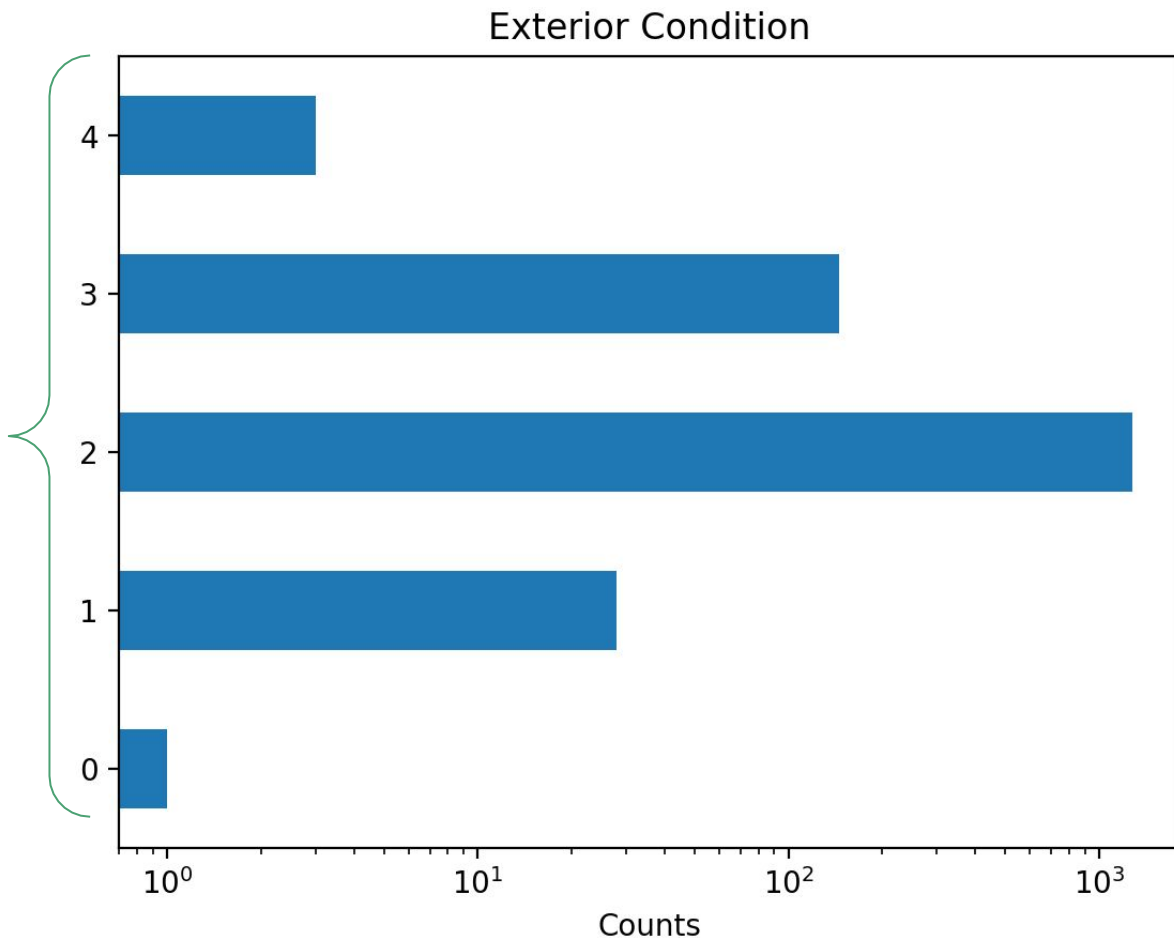
Кейде бізге қажет:



Неліктен
инженерлік
мүмкіндіктер?

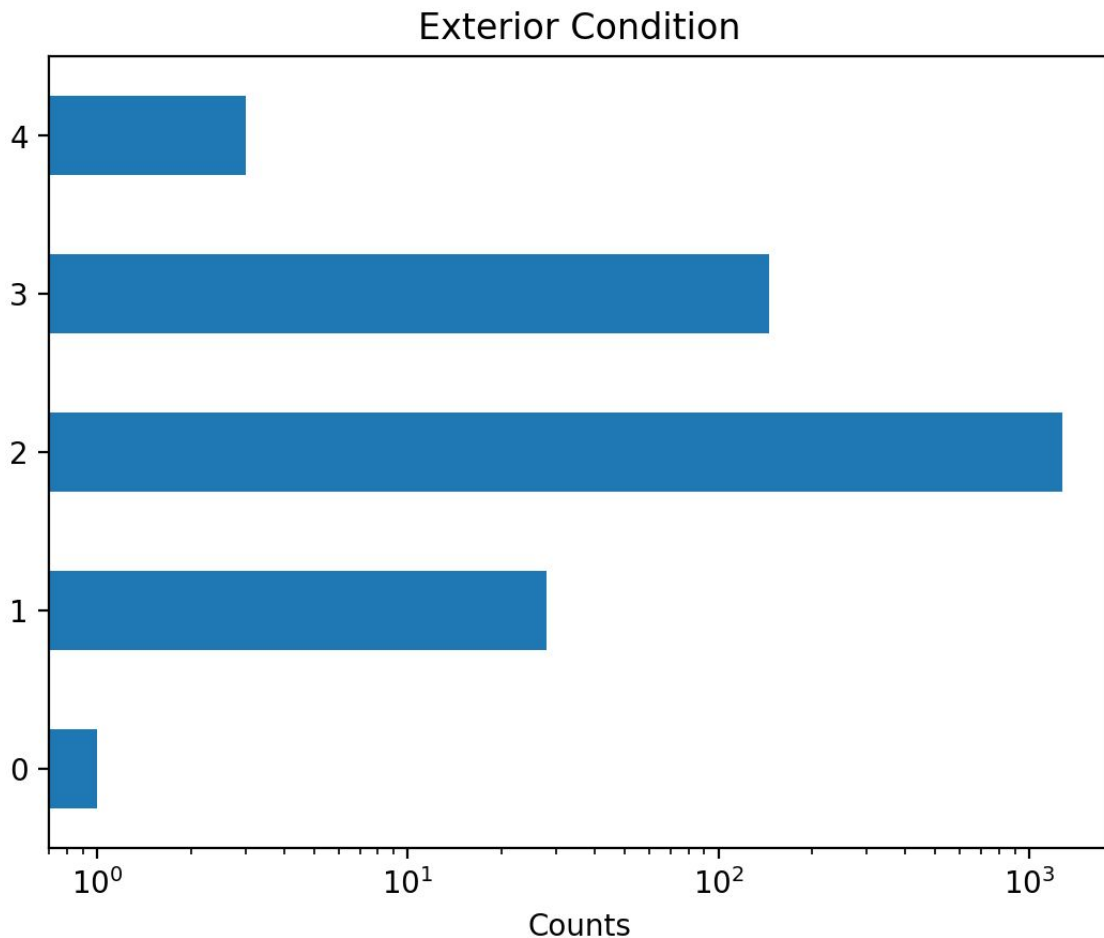
Кейде бізге қажет:

Сандар!



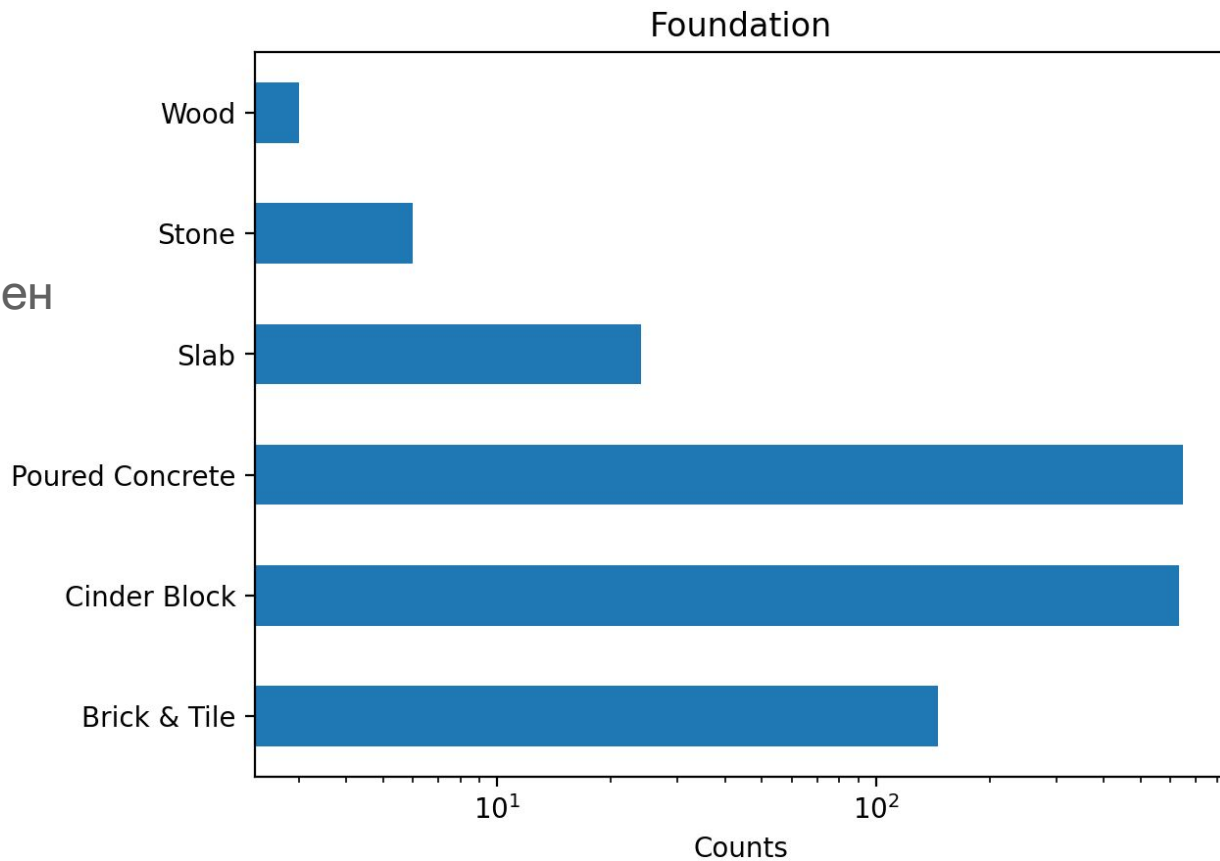
Реттік кодтау

- Егер санаттардың бір-бірінен әртүрлі «қашықтықтары» болса, реттік кодтауды орындаңыз.
- **мысалы**
- тым кішкентай, жақсы сәйкес келеді, тым үлкен
- Толық келіспеймін, келіспеймін, бейтарап, келісемін, қатты келісемін



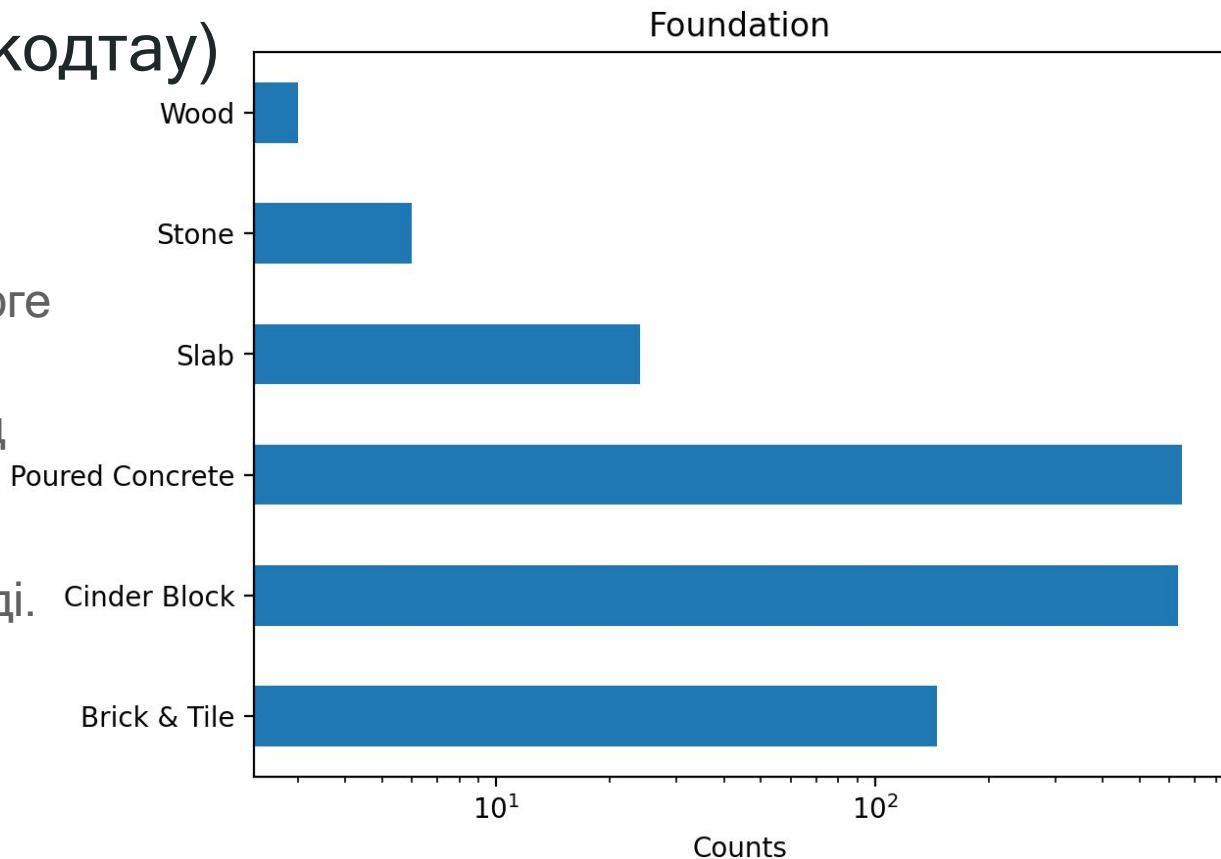
Реттік емес санаттар

- Кейде санаттар негізінен тәуелсіз болады.



One Hot Encoding (кодтау)

- С санаттарын елестетіңіз.
- С-1 екілік мүмкіндіктерге бөліңіз
- Әрбір мүмкіндік үлгінің белгілі бір санатқа жататынын немесе жатпайтынын көрсетеді.
- Егер барлық мүмкіндіктер False болса, бұл үлгінің Cth санатына жататынын білдіреді.



One Hot Encoding

	<i>is_wood</i>	<i>is_stone</i>	<i>is_slab</i>	<i>is_poured</i>	<i>is_cinder</i>
«Ағаш»	1	0	0	0	0
«Тақта»	0	0	1	0	0
«Кірпіш және плитка»	0	0	0	0	0

One Hot Encoding бір ескертуі

- Деректер жиынының өлшемін арттырады: $n \times 1 \rightarrow n \times C$
- Үлкен C үшін неғұрлым жетілдірілген мүмкіндікті инженериямен біріктіру қажет.
- Үлкен C мысалдары:
 - Ұсынатын жүйелер: әрбір пайдаланушы мен элементті бір реттік кодтау.
 - Мәтін: әрбір бірегей сөзді немесе ішкі сөзді бір реттік кодтау.
- Үлкен C шешімдері :
 - Енгізу: барлық санаттарды төмен өлшемді векторлық кеңістікке ендіру.
 - Әрбір санат $O(100)$ өлшемді вектормен салыстырылады.
 - Векторлар модель параметрлері болып табылады. Енгізілетін үлгі параметрлерін үйреніңіз осы векторлық кеңістікте бір-біріне жақын мағыналық ұқсас категориялар.
 - Хэшинг: санаттар кездейсоқ екілік мүмкіндіктерге хэштеледі.
 - Санаттар санынан аз мүмкіндіктерге хэш.
 - Кішірек мүмкіндік кеңістігі үшін төмендетілген дәлдікті (хэш соқтығыстарына байланысты) ауыстырыңыз.

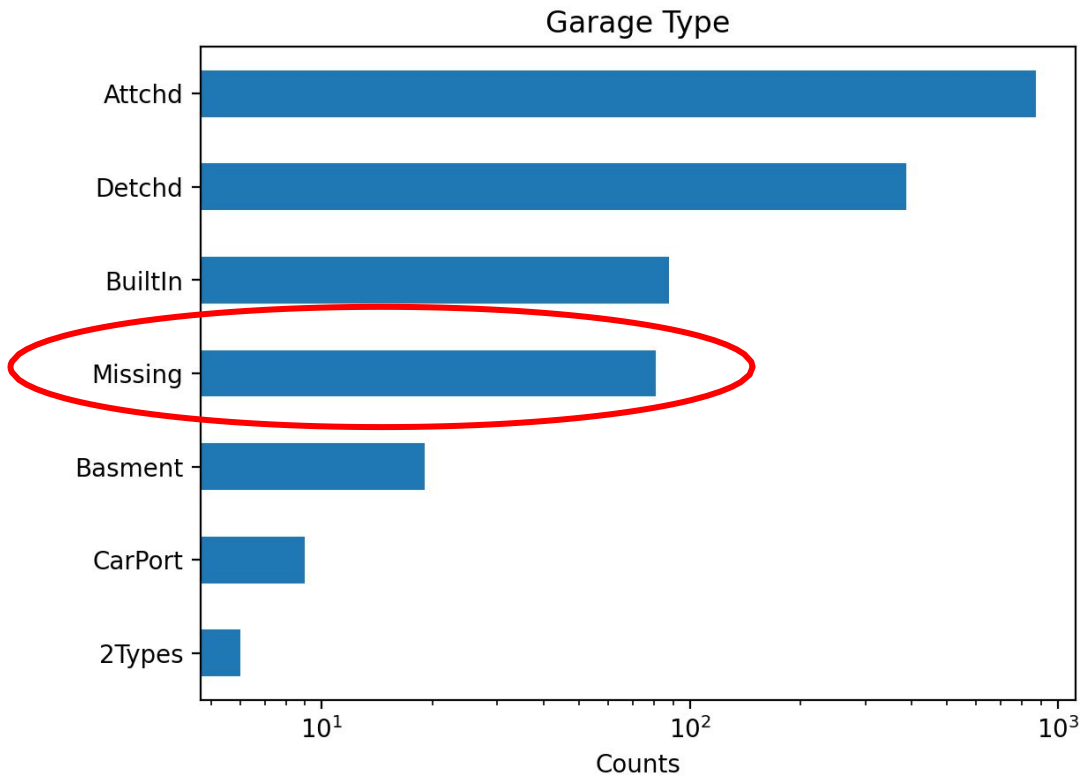
инженериясы – масштабтау мен кодтаудан тыс

- Ағындар -> Мүмкіндіктер
- “Соңғы айдағы орташа тапсырыс құны”
 - «Пернелерді басу арасындағы уақыттың стандартты ауытқуы»
 - «Соңғы аптада тұтынушының осы бейне санатын көрген жалпы минуттары»
- Суреттер
 - Модельдер -> модельдер
 - Сурет -> «Көліктің алдындағы жаяу жүргінші» -> «тежеу керек»
 - Сурет -> «Көліктің алдындағы жаяу жүргінші» -> «тежеу керек
 - Шикі пикселдермен беру және мүмкіндіктерді үйреніңіз
- Мәтін
 - Аннотация
 - Сөйлеуді тегтеу бөлігі
 - Сезімдерді талдау
 - Атаулы нысанды тану

Санат жоқ

Ең қарапайым шешім:

- Жаңа «жоғалған» санатты жасаңыз, содан кейін кодтаңыз.



Функция инженериясы - Сандық деректер жоқ

- Жетіспейтін деректерді өңдеу әдетте «импутация» деп аталады.

Қарапайым шешімдер:

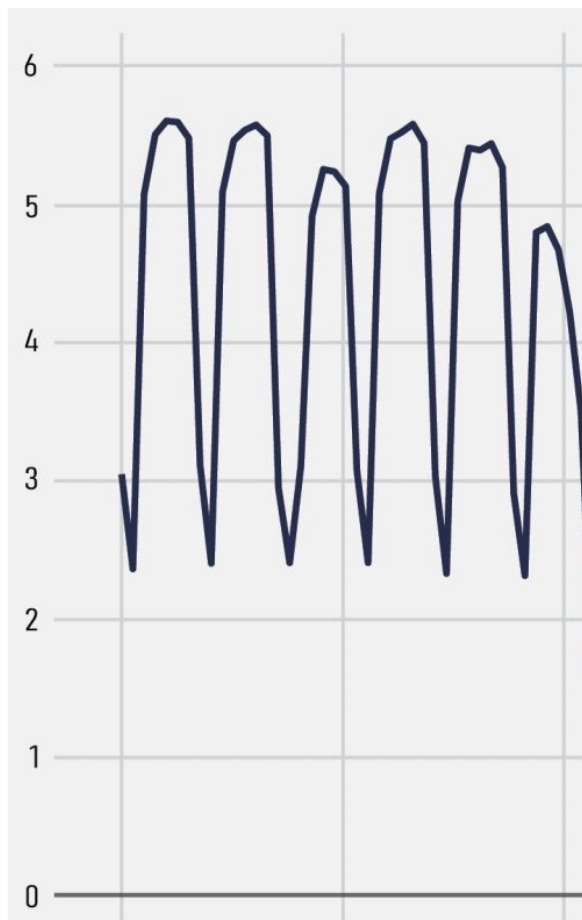
Жетіспейтін мәндерді орташа мәнмен толтырыңыз

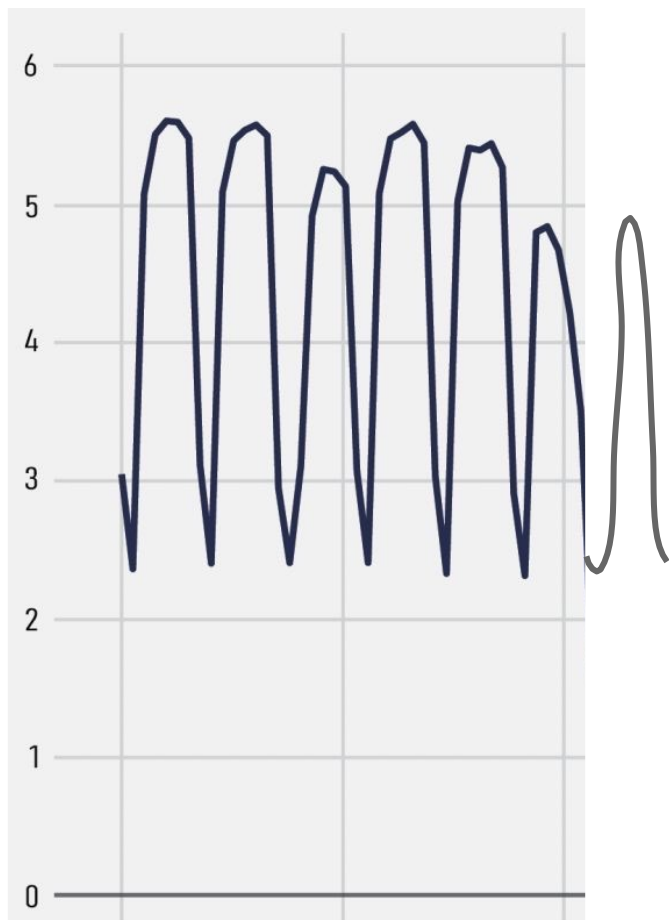
- 0-мен жетіспейтін мәндерді толтырыңыз және деректер жоқ/жоқ болған кезде 1/0 болатын бөлек «индикатор» екілік мүмкіндігін жасаңыз.

Абайлаңыз!

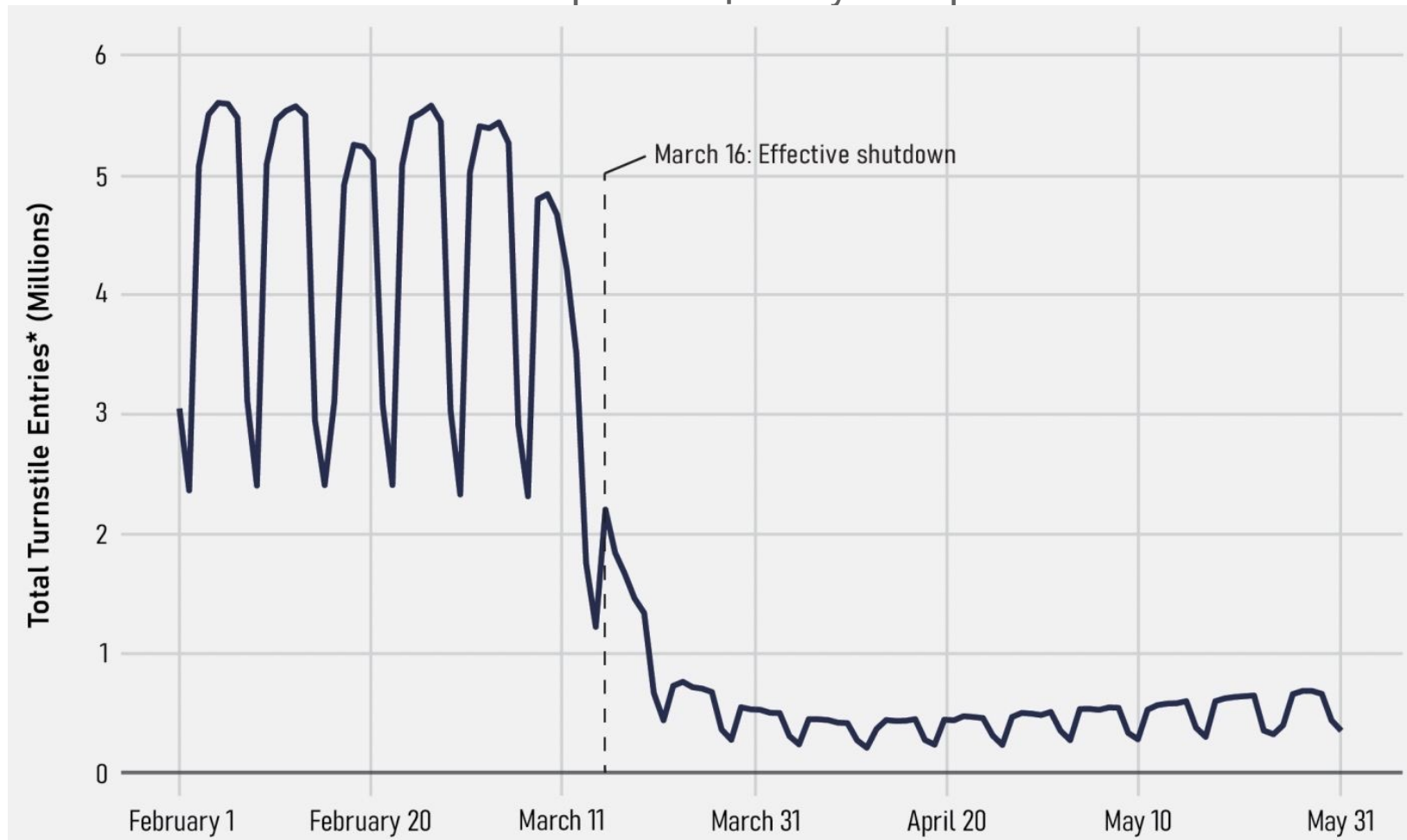
- Деректер кездейсоқ жоғалып кетпеуі мүмкін.
- Жетіспейтін деректердің өзі құнды сигнал болуы мүмкін.
- Деректер жиынтығынан жетіспейтін деректерді жай ғана жоймаңыз.

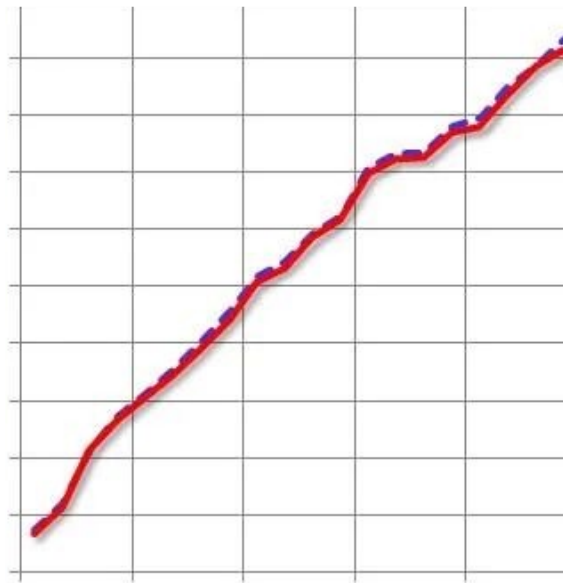
Үлгіні таңдау

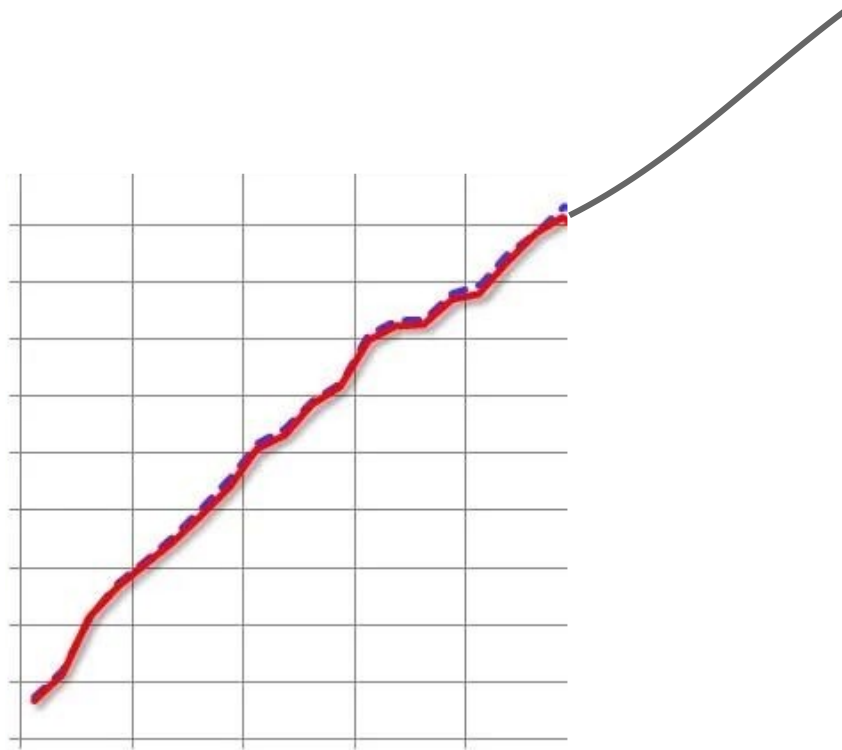




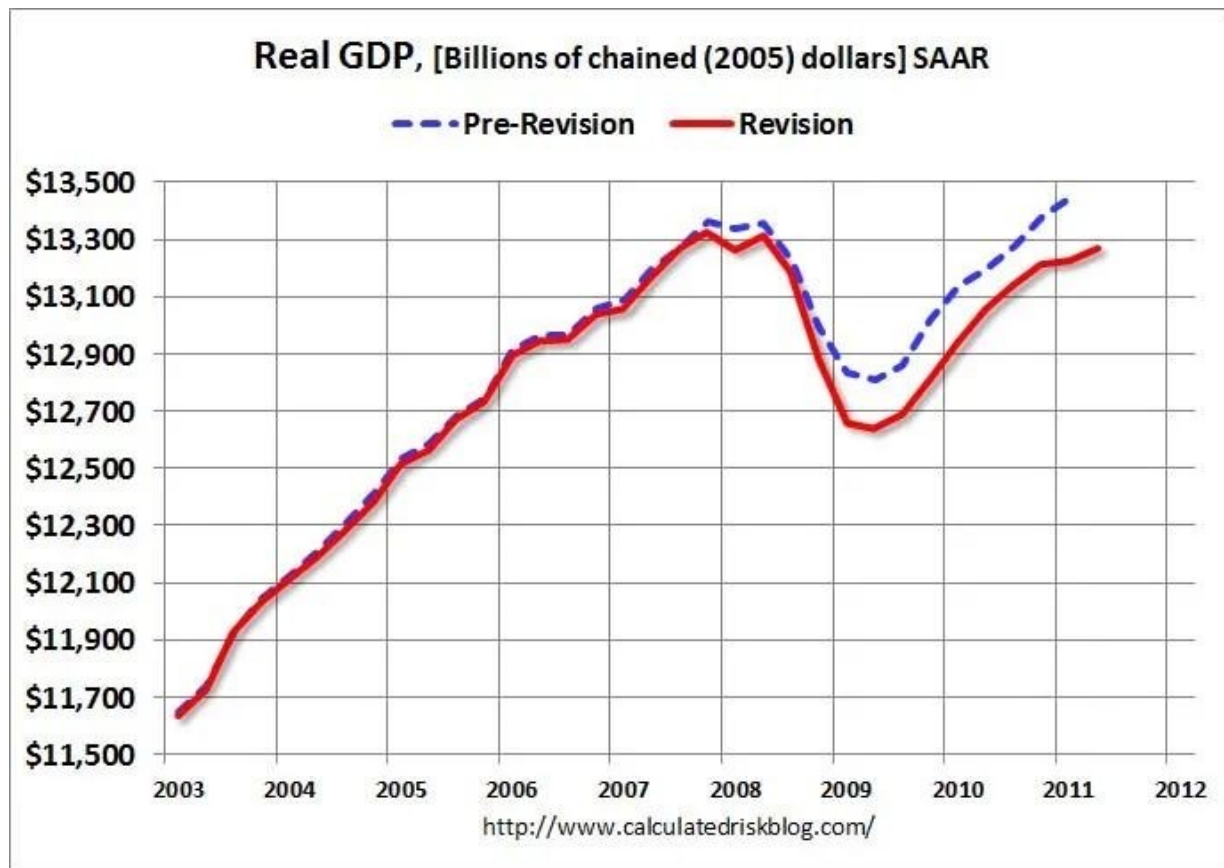
MTA метросының жолаушылар







АҚШ ЖІӨ және Ұлы рецессия



Тақырып қандай?

Жаңа деректер модель үйренген деректерден өзгеше

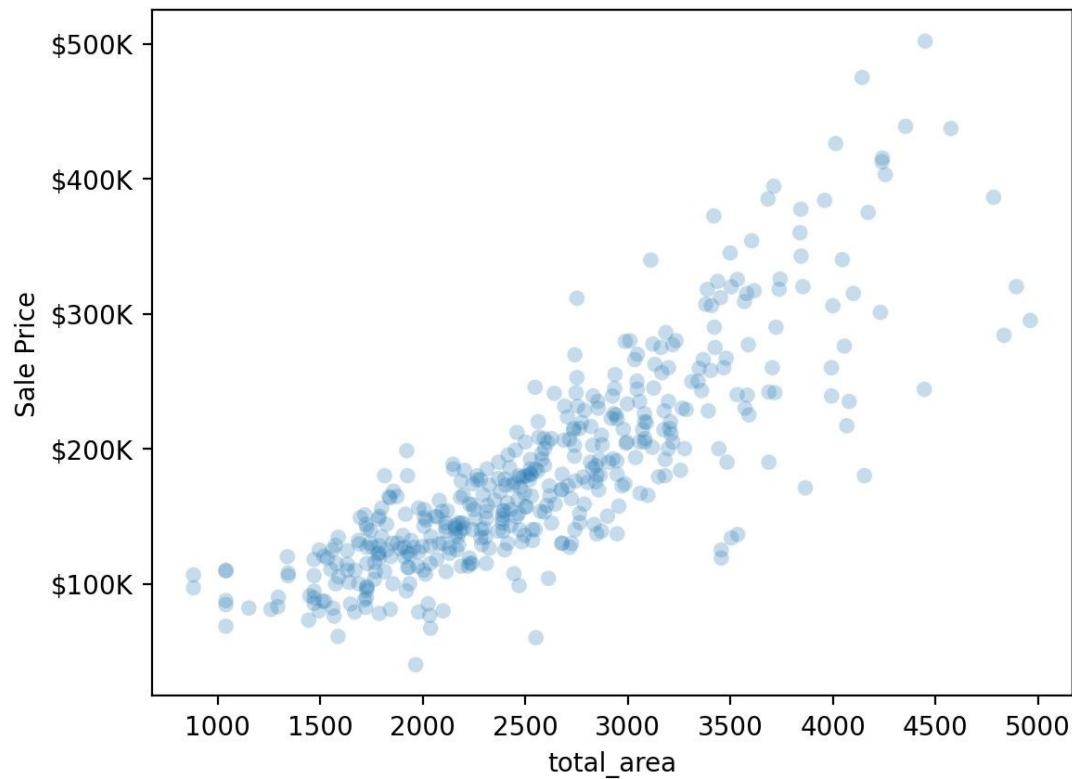
Модельді таңдау мақсаты

- Біздің өнімділік талаптарына сәйкес келетін «модель» жасаңыз
- Біздің өнімділік өлшеміміз «нақты әлемде» орындалатынына сенімді болыңыз
- Біз модельдің өндірісте қалай қолданылатынын мүмкіндігінше жақсырақ болжағымыз келеді.

Үлгіні таңдау

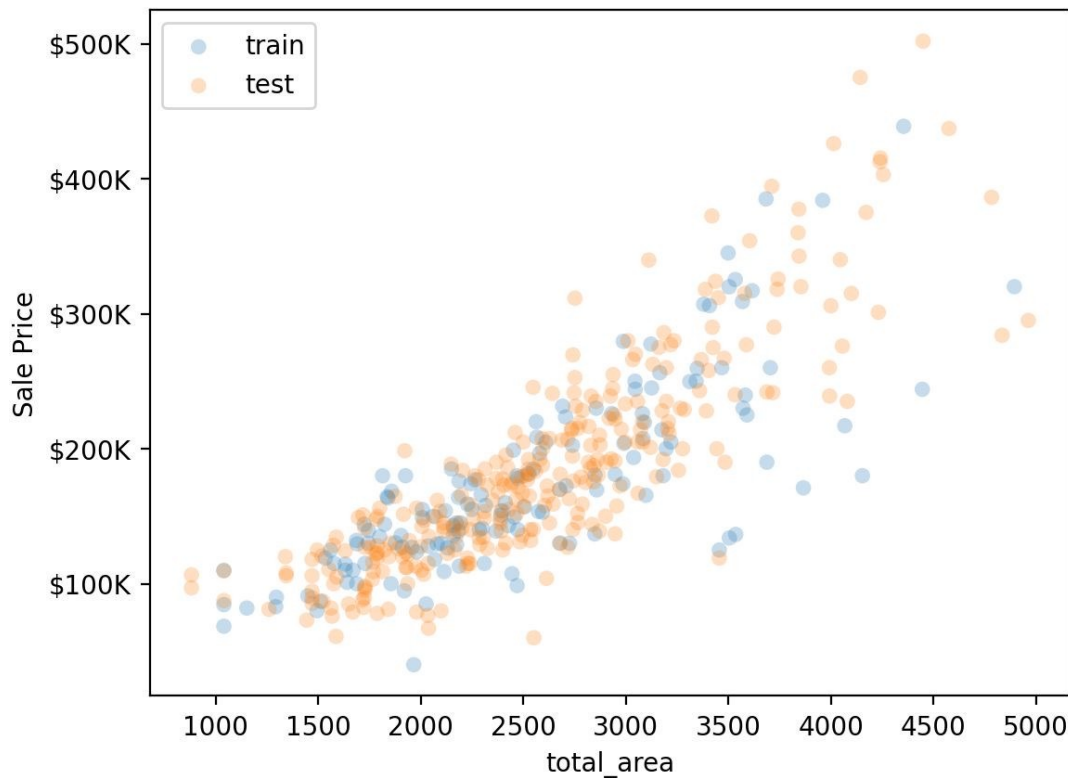
- Біздің өнімділік талаптарына сәйкес келетін «модель» жасаңыз
- Біздің өнімділік өлшеміміз «нақты әлемде» орындалатынына сенімді болыңыз
- Біз модельдің өндірісте қалай қолданылатынын мүмкіндігінше жақсырақ болжағымыз келеді.

Holdout жиынтығы, Train/Test Split



Holdout жиынтығы, Train/Test Split

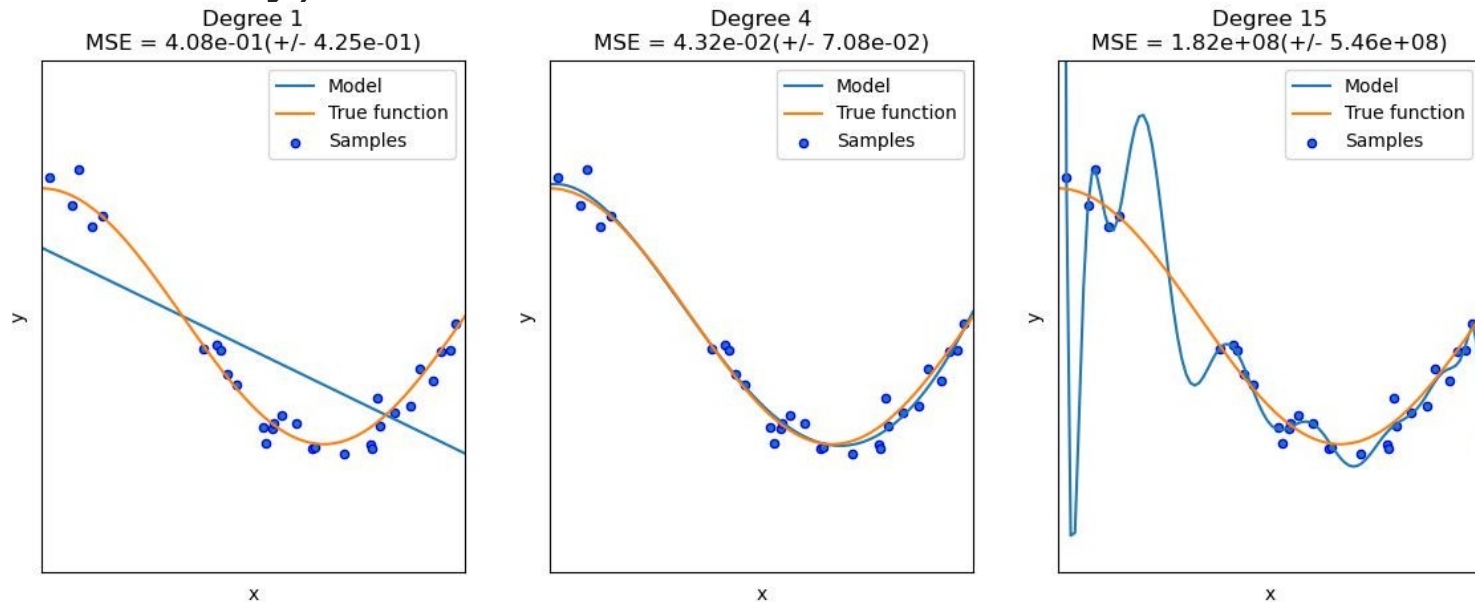
- Деректерді оқу және сынақ деректер жиындарына кездейсоқ бөліңіз.
- Модельді тек жаттығу жиынтығына үйретіңіз.
- Модельді сынақ жинағы арқылы бағалаңыз.
- Егер деректер IID болса, кездейсоқ іріктеу бізге модельдің «шын» өнімділігінің объективті бағасын бере алады.



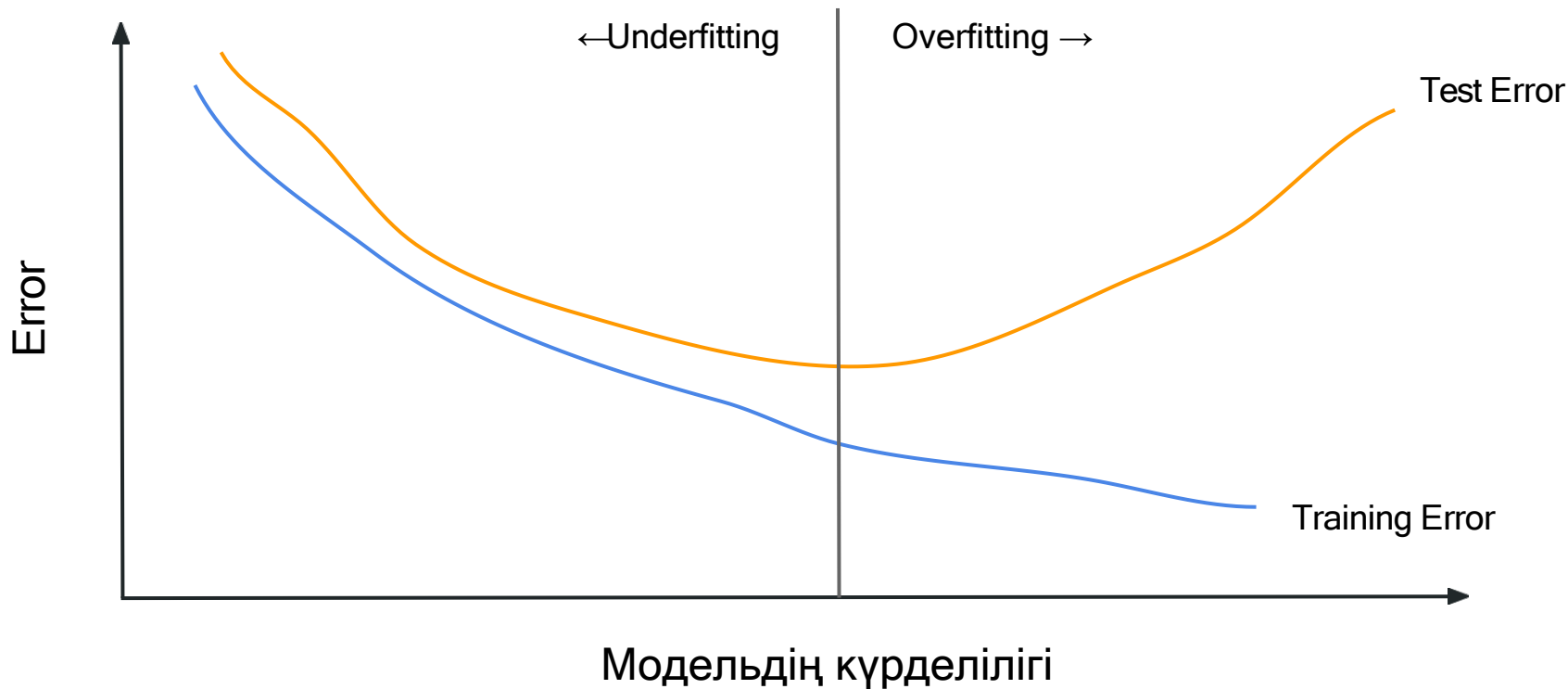
“Төрт параметрмен мен
пілді сыйдыра аламын, ал
бес параметрмен оны діңін
айналдыра аламын.”

— John Von Neumann

Overfitting and Underfitting (Артық қондыру және жарамсыз ету)



Overfitting and Underfitting: Үлгінің күрделілігін өзгерту



Overfitting and Underfitting: Регуляризация

$$\mathcal{L} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$\mathcal{L} = \sum_{i=1}^n \left(y_i - \vec{\mathbf{X}}_i \cdot \vec{\beta} \right)^2$$

Overfitting and Underfitting: Регуляризация

$$\mathcal{L} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$\mathcal{L} = \sum_{i=1}^n \left(y_i - \vec{\mathbf{X}}_i \cdot \vec{\beta} \right)^2$$

$$\mathcal{L} = \sum_{i=1}^n \left(y_i - \vec{\mathbf{X}}_i \cdot \vec{\beta} \right)^2$$

L2 Регуляризация /
Жоталардың
регрессиясы




$$+ \lambda_2 \sum_{j=1}^p \beta_j^2$$

Overfitting and Underfitting: Регуляризация

$$\mathcal{L} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$\mathcal{L} = \sum_{i=1}^n \left(y_i - \vec{\mathbf{X}}_i \cdot \vec{\beta} \right)^2$$

L1 Regularization / Lasso



$$\mathcal{L} = \sum_{i=1}^n \left(y_i - \vec{\mathbf{X}}_i \cdot \vec{\beta} \right)^2 + \lambda_1 \sum_{j=1}^p |\beta_j|$$

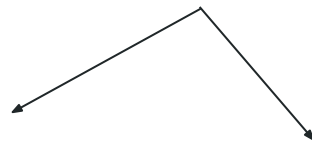
Overfitting and Underfitting: Регуляризация

$$\mathcal{L} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$\mathcal{L} = \sum_{i=1}^n \left(y_i - \vec{\mathbf{X}}_i \cdot \vec{\beta} \right)^2$$

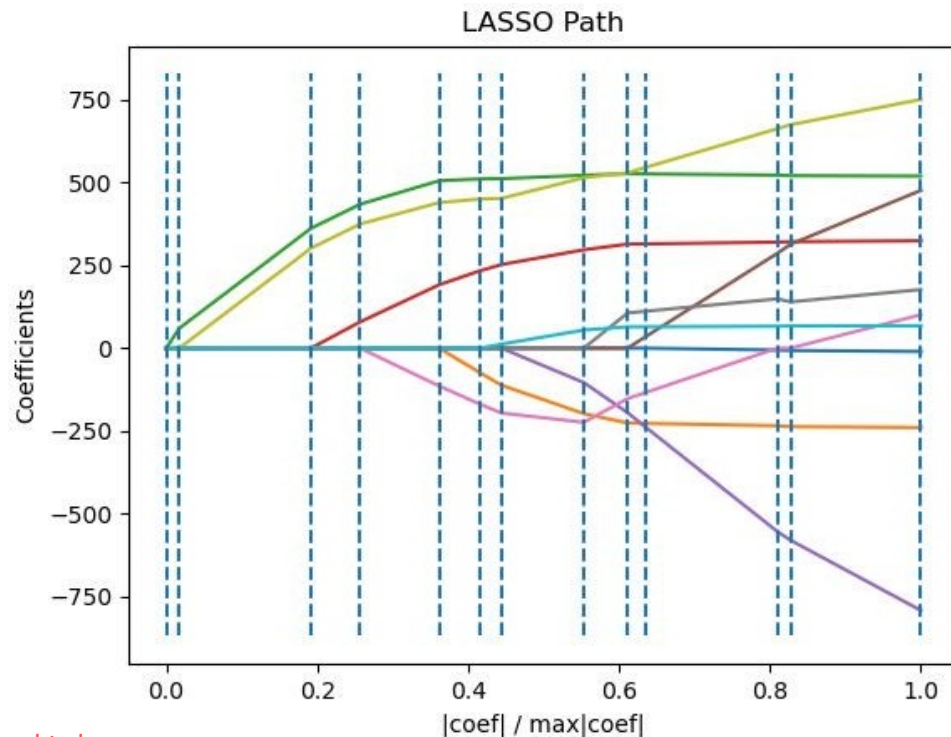
$$\mathcal{L} = \sum_{i=1}^n \left(y_i - \vec{\mathbf{X}}_i \cdot \vec{\beta} \right)^2 + \lambda_1 \sum_{j=1}^p |\beta_j| + \lambda_2 \sum_{j=1}^p \beta_j^2$$

Elastic Net



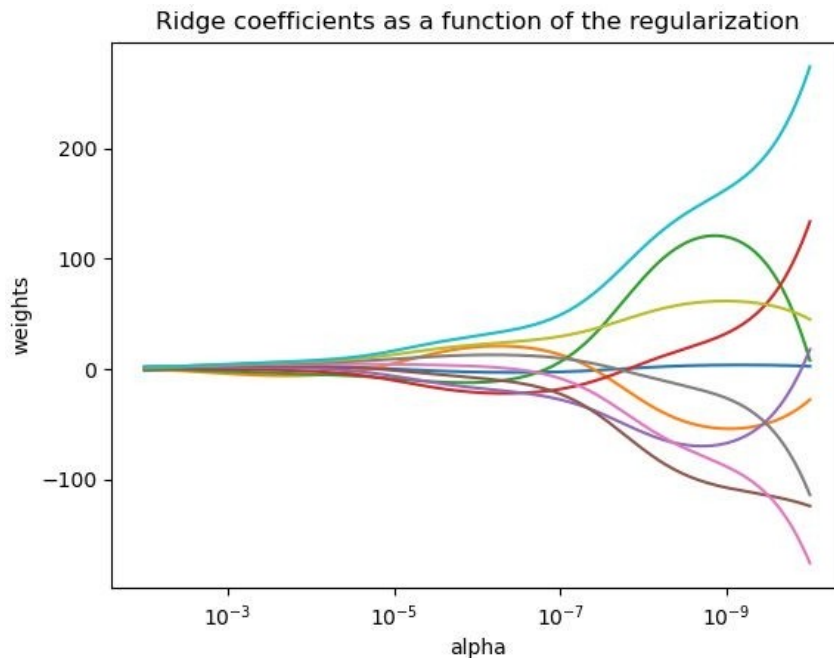
Overfitting and Underfitting: Регуляризация

- L1 Регуляризация «сиректікті» тудыруы мүмкін.
- Реттеу күшін арттырған сайын мүмкіндік коэффициенттері нөлге дейін төмендейді.
- Мұны мүмкіндікті таңдау үшін пайдалануға болады.
- Егер мүмкіндіктер корреляцияланса, біреуі 0-ге дейін төмендеп, екіншісі үлкен болып қалуы мүмкін.

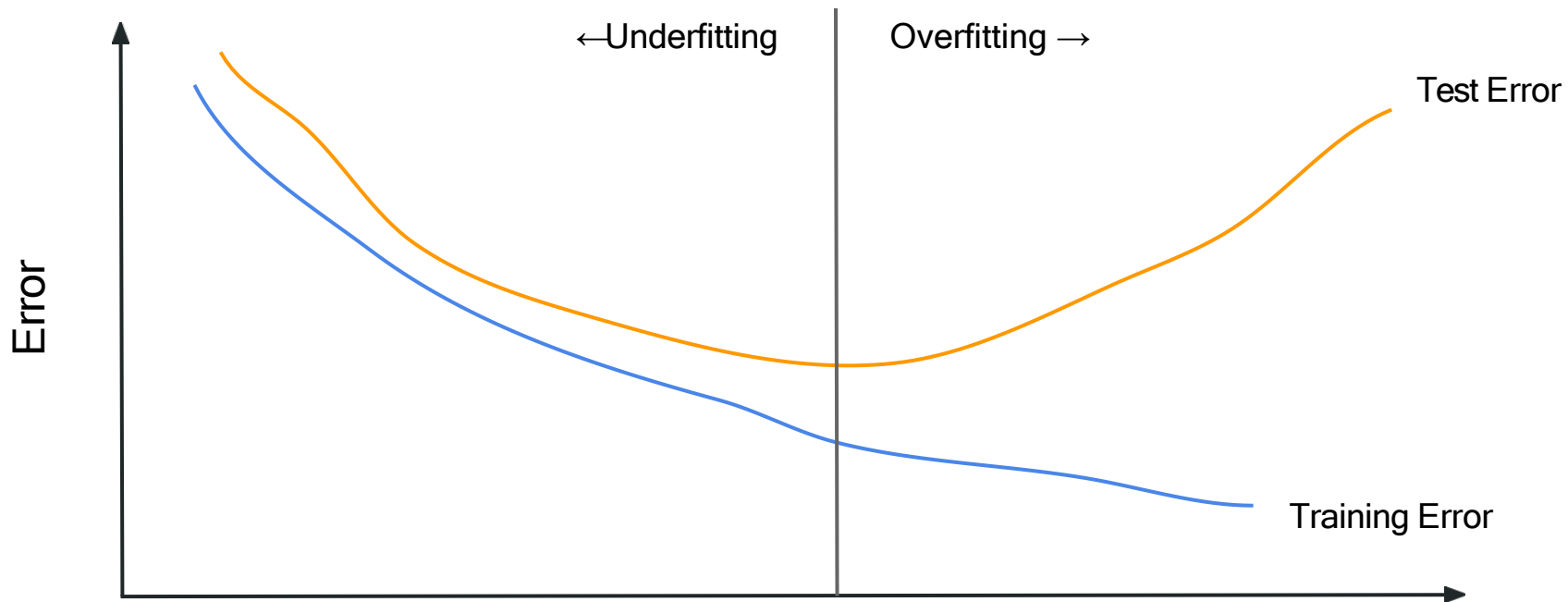


Overfitting and Underfitting: Регуляризация

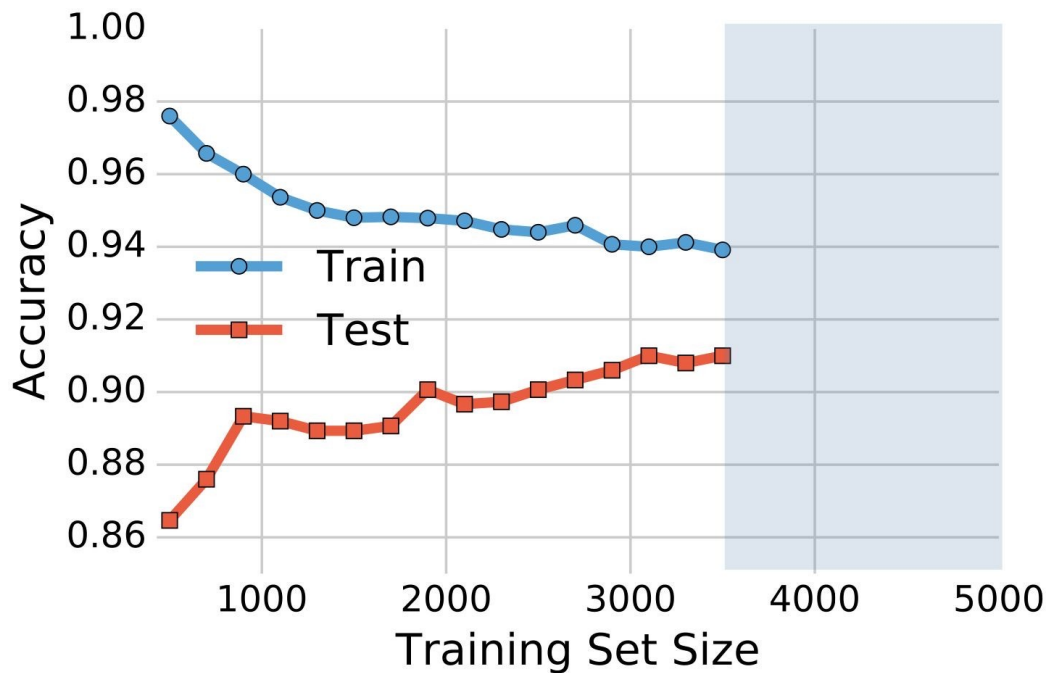
- L2 Регуляризация барлық коэффициенттерді бірге азайтады.
- Корреляциялық мүмкіндіктер бірге төмендейді.
- Сиректікті тудырмайды.
- Регуляризация - бұл «гиперпараметрдің» түрі.
- Гиперпараметр - оңтайландыру процесінің бөлігі ретінде үйренбейтін үлгі параметрі.
- https://scikit-learn.org/stable/auto_examples/linear_model/plot_ridge_path.html



Overfitting and Underfitting: Регуляризация



Overfitting and Underfitting: Қосымша деректерді пайдаланыңыз



Raschka (2018)

Figure 4: Learning curves of softmax classifiers fit to MNIST. <https://arxiv.org/abs/1811.12808>