အခန်း (၂): Regression methods

၁။ Regression Analysis ဆိုသည်မှာ ကြိုတင်စုဆောင်းထားသည့် အချက်အလက်များကို အခြေခံ၍ ရလဒ်နှင့် အချက်အလက်များ၏ ယေဘူယျ ဆက်သွယ်ချက်ကို ရှာဖွေခြင်း ဖြစ်သည်။ ထို့နောက် အဆိုပါ ယေဘူယျ ဆက်သွယ်ချက်ကို အသုံးပြု၍ အချက်အလက် အသစ် အတွက် ရလဒ်ကို ခန့်မှန်းခြင်း ဖြစ်သည်။ Regression Analysis ကို လေ့လာရာတွင် သိထားသင့်သည့် အသုံးအနှုန်း (၅) ခုနှင့် စတင် မိတ်ဆက် ပေးလိုပါသည်။

(က) Training data ဆိုသည်မှာ ကြိုတင် စုဆောင်းထားသည့် အချက်အလက်များကို ဆိုလိုသည်။ Training data တခုချင်းစီတွင် ရလဒ်နှင့် အဆိုပါ ရလဒ်ကို သက်ရောက်စေနိုင်သည့် အချက်အလက်များ ပါဝင်ရမည်။

(ခ) Target variable ဆိုသည်မှာ ရလဒ်ကို ရည်ညွှန်းခြင်း ဖြစ်သည်။ ဥပမာ -- အိမ်ဈေး ခန့်မှန်းသည့် ပုစ္ဆာတခုအတွက် အိမ်ဈေးသည် Target(ရလဒ်) ဖြစ်ပြီး အဆိုပါ ဈေးနှုန်းသည် တည်နေရာ၊ အကျယ်အဝန်း၊ စသည့် အခြားအချက်အလက်များအပေါ်တွင် မူတည် သည်။ သို့ဖြစ်ရာ Target variable ကို dependent Variable (မှီခို ကိန်းရှင်) အဖြစ်လည်း ဖလှယ် သုံးစွဲလေ့ ရှိသည်။

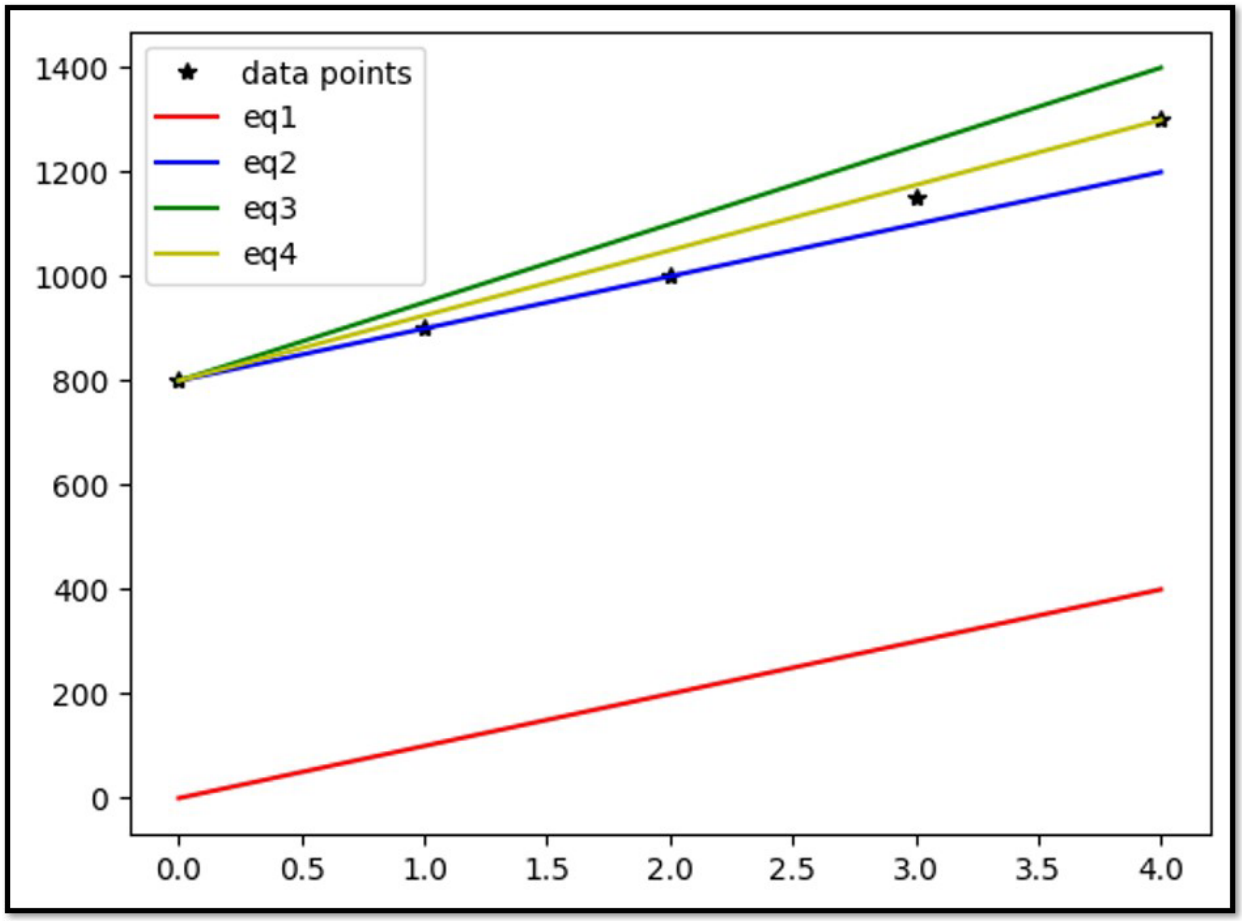
(ဂ) Independent Variables ဆိုသည်မှာ အထက်ပါ ရလဒ်ကို အကျိုးသက်ရောက်စေနိုင်သည့် အချက်အလက်များကို ရည်ညွှန်းသည်။ အထက်ပါ အိမ်ဈေး ခန့်မှန်းသည့် ပုစ္ဆာတွင် တည်နေရာ၊ အိမ်၏ အကျယ်အဝန်း၊ အိပ်ခန်း အရေအတွက် စသည့် အချက်အလက်များသည် Independent Variable များ ဖြစ်ကြသည်။ သတိပြုရန်မှာ ရလဒ် (ဥပမာ -အိမ်ဈေး) သည် ၁ ခု ထက်မကသည့် အချက်အလက်များအပေါ်တွင် မူတည်သည်။

(ဃ) Parameters ဆိုသည်မှာ ရလဒ်နှင့် အချက်အလက်များ၏ ယေဘူယျ ဆက်သွယ်ချက်ကို ဖော်ပြနိုင်သည့် ကိန်းသေများကို ရည်ညွှန်းသည်။ ဥပမာ ရုန်ကုန်တွင် တိုက်ခန်း ရောင်းချရာ၌ အိမ်အကျယ်အဝန်း ၁ ပေပတ်လည်ကို ၁ သိန်း/၁ သိန်းခွဲ စသည်ဖြင့် သတ်မှတ် ရောင်းချသည်ကို ကြုံဖူးကြပါလိမ့်မည်။ အဆိုပါ ဆက်သွယ်ချက်ကို သချိၤ ညီမျခြင်းဖြင့် ဖော်ပြမည်ဆိုပါက အိမ်ဈေး = ကိန်းသေ x အိမ်အကျယ်အဝန်း ဖော်ပြနိုင်ပြီး အဆိုပါ ကိန်းသေကို Parameters ဟု ခေါ်ဆိုခြင်း ဖြစ်သည်။

(င) Residuals ဆိုသည်မှာ **machine learning method** မှ ခန့်မှန်းလိုက်သော ရလဒ်နှင့် မူလ ရလဒ်၏ ခြားနားချက် ဖြစ်သည်။

Linear Regression

၂။ Linear Regression Analysis တွင် ရလဒ်နှင့် အချက်အလက်များသည် မျဉ်းဖြောင်း တကြောင်း ထဲဖြင့် ဆက်သွယ်ထားသည်ဟူသော ယူဆချက်ကို အခြေခံ၍ သုံးသပ်ခြင်းဖြစ်သည်။ အထက်ပါ အိမ်စျေး ခန့်မှန်းသည့် ပုစ္ဆာကို ပြန်လည် ညွှန်းဆိုရပါက အိမ် ၁ ပေပတ်လည် တက်သွားတိုင်း အိမ်စျေးမှာ ၁ သိန်း ခွဲတက်သွားမည် ဖြစ်ရာ တက်သွားသည့်နှုန်းမှာ တပြေးညီထဲ ဖြေနေမည်ဟု ယူဆထားခြင်း ဖြစ်သည်။ လက်တွေ့တွင်မူ အိမ်အကျယ်အဝန်း ကြီးလာသည်နှင့်အမျှ စျေးလျော့ပေးခြင်းများလည်း ရှိနိုင်ရာ တပြေးညီ ဖြစ်နေသည်ဟုသော ယူဆချက်သည် အမြဲ မှန်ကန်မှု မရှိနိုင်ပါ။



ပုံ (၁) Simple Linear Regression ဥပမာများ

Simple Linear Regression

၃။ Simple Linear Regression တွင် ရလဒ်သည် အချက်အလက်တခုထဲ အပေါ်သာ မူတည်သည့် ဟူသော ယူဆချက်ကို အသုံးပြု၍ ရလဒ် (y) နှင့် အချက်အလက်တခု (x) ၏ ဆက်သွယ်ချက်ကို အောက်ပါ သင်္ချာ ညီမျှခြင်းဖြင့် ဖော်ပြလေ့ ရှိသည်။

$$y = \omega_1 x + b \qquad (၂.၁)$$

၄။ အထက်ပါ ညီမျှခြင်းကို ပုံဆွဲကြည့်မည်ဆိုပါက မျဉ်းဖြောင့်တကြောင်းကို ရရှိမည် ဖြစ်သည်။ Parameter ω_1 နှင့် b ၏ နေရာတွင် ကိန်းဂဏန်းအမျိုးမျိုးကို အစားထိုးခြင်းအားဖြင့် ပုံ ၁ တွင် ပြထားသည့် မတူ ညီသည့် မျဉ်းဖြောင့်များ ကို ရရှိပါသည်။

၅။ Simple Linear Regression ၏ ရည်ရွယ်ချက်မှာ **ခန့်မှန်းသည့် အဖြေနှင့်ရလဒ်အမှန်တို့ အကြား ကွာဟမှု** အနည်းဆုံး ဖြစ်စေသည့် Parameter (ကိန်းသေ) များကို ရှာဖွေရန် ဖြစ်သည်။

Cost Function

၆။ Simple Linear Regression model တခု၏ ခန့်မှန်းသည့် အဖြေ $(\widetilde{y_i})$ နှင့်ရလဒ်အမှန် (y_i) တို့ အကြား ပျမ်းမျ ကွာဟမှုကို အောက်ပါ သင်္ချာ ညီမျှခြင်းသုံး၍ ဖော်ပြနိုင်ပါသည်။ ခန့်မှန်းသည့် အဖြေ $(\widetilde{y_i})$ ကို ညီမျှခြင်း ၂.၁ ကို အသုံးပြု၍ တွက်ချက်နိုင်သည်။

$$\in = \frac{1}{N}\sum_{i=1}^{N}(\tilde{y}_i - y_i) \qquad (၂.၂)$$

၇။ အထက်ပါ Cost Function ၏ အနည်းဆုံး တန်ဖိုးဖြစ်စေသည့် Parameter ω_1 နှင့် b ကို ရှာဖွေရန်အတွက် Gradient Descent method ကို အသုံးပြုလေ့ ရှိသည်။

Multiple Linear Regression

၈။ Multiple Linear Regression တွင် ရလဒ်သည် ၁ ခု ထက်မကသည့် အချက်အလက်များ အပေါ်တွင် မူတည်သည်ဟုသော ယူဆချက်ကို အသုံးပြုထားရာ ရလဒ် (y) နှင့် အချက်အလက်များ ၏ ဆက်သွယ်ချက်ကို အောက်ပါ သင်္ချာ ညီမျှခြင်းဖြင့် ဖော်ပြလေ့ ရှိသည်။

$$y = \omega_1 x_1 + \omega_2 x_2 + \cdots + b \qquad (၂.၃)$$

၉။ Simple Linear Regression ကဲ့သို့ပင် Multiple Linear Regression ၏ ရည်ရွယ်ချက်မှာ **ခန့်မှန်းသည့် အဖြေနှင့်ရလဒ်အမှန်တို့ အကြား ကွာဟမှု** အနည်းဆုံး ဖြစ်စေသည့် Parameter (ကိန်းသေ) များကို ရှာဖွေရန် ဖြစ်သည်။

Feature Scaling

၁၀။ Simple Linear Regression နှင့် Multiple Linear Regression တို့၏ အဓိက ကွာခြားချက်မှာ အချက် အလက် ၁ ခု ထဲကို အသုံးပြုခြင်းနှင့် အချက်အလက်များကို အသုံးပြုခြင်း ဖြစ်သည်။ ၁ ခု ထက်ပိုသော အချက်အလက်များကို အသုံးပြုရာတွင် သတိပြုရမည့် အချက်မှာ အချက်အလက် တခု နှင့် တခုကြား ပမာဏ မတူညီခြင်း ဖြစ်သည်။

၁၁။ ဥပမာ တိုက်ခန်း တခန်း၏ ဈေးနှုန်းသည် အကျယ်အဝန်းနှင့် အိပ်ခန်း အရေအတွက်ပေါ် မူတည်သည်ဟူသောအချက်ကို အခြေခံ၍ ယေဘူယျ ဆက်သွယ်ချက်ရှာဖွေကြမည် ဆိုပါစို့။ အိမ်တအိမ်၏ အကျယ်အဝန်းမှာ ပေ ၅၀၀ မှ ပေ ၁၀၀၀ ကျော် ပတ်လည်ထိ ရှိနိုင်သော်လည်း အိပ်ခန်း အရေအတွက်မှာမူ ၁၀ ခန်းထက် မကျော်နိုင်ပါ။ မတူညီသည့် တန်ဖိုး ၂ ခုကို သင်္ချာ ညီမျှခြင်းတကြောင်းထဲကို အသုံးပြု၍ ဖြေရှင်းမည်ဆိုပါက တိကျ မှန်ကန်နိုင်ခြင်း မရှိပါ။ သို့ဖြစ်ရာ Multiple Linear Regression ကို အသုံးပြုမည်ဆိုပါက အချက်အလက်များကို တူညီသည့် တန်ဖိုး တခုထဲတွင် ရှိစေရန် ညှိပေးရမည် ဖြစ်သည်။ ထိုသို့ ပြုလုပ်ခြင်းကို Feature Scaling ဟု ခေါ်ဆိုပါသည်။ ဤစာအုပ်တွင် အသုံးများသည့် Feature Scaling နည်းလမ်း ၂ သွယ်ကို ဖော်ပြသွားပါမည်။

Min-Max Scaling

၁၂။ Min-Max Scaling ဆိုသည်မှာ အချက်အလက် တခုချင်းစီ၏ တန်ဖိုးကို သုညနှင့် တစ်ကြား ရောက်အောင် ညှိယူခြင်း ဖြစ်သည်။ သချိုာအားဖြင့် အောက်ပါ အတိုင်း ဖော်ပြနိုင်သည်။

$$x_{scale} = \frac{x - x\_min}{x\_\max - x\_min} \qquad (၂.၄)$$

Z-score Standardization

၁၃။ Z-score Standardization ဆိုသည်မှာ အချက်အလက် တခုချင်းစီ၏ ပျမ်းမျ တန်ဖိုးကို သုည ဖြစ်စေပြီး ပျမ်းမျတန်ဖိုးမှ ခြားနားချက်ကို ၁ အတွင်း ရရှိအောင် ပြုလုပ်ခြင်း ဖြစ်သည်။ သချိုာအားဖြင့် အောက်ပါ အတိုင်း ဖော်ပြနိုင်သည်။

$$x_{scale} = \frac{x - \mu}{\sigma} \qquad (၂.၅)$$

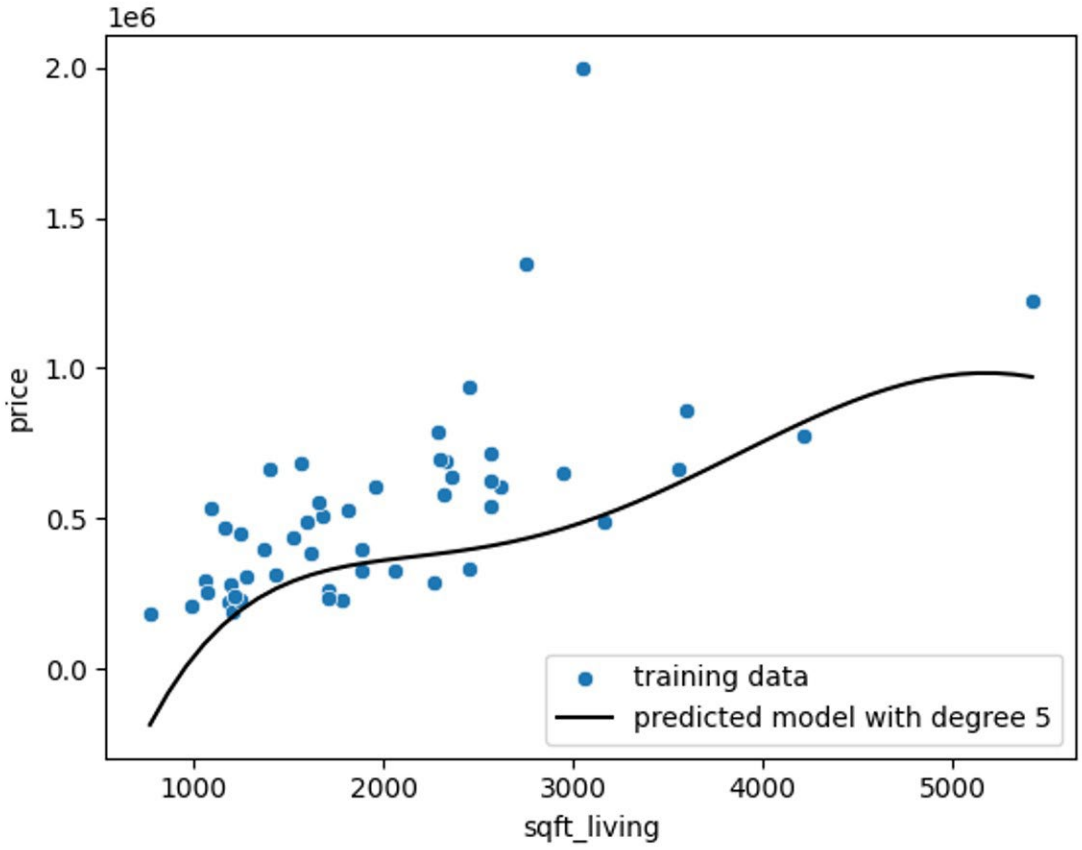
၁၄။ Linear Regression ကို အသုံးပြုရာတွင် အထူးသတိပြုရမည့် အချက်မှာ ရလဒ်နှင့် အချက် အလက်များအကြား ဆက်သွယ်ချက်မှာ Linear ဖြစ်နေရမည် ဖြစ်သည်။ တနည်းအားဖြင့် ရလဒ်နှင့် အချက်အလက်များ၏ ဆက်သွယ်ချက်ကို ပုံဖြင့် ဖော်ပြမည်ဆိုပါက မျဉ်းဖြောင့်တကြောင်းဖြင့် ဖော်ပြနိုင် ရမည် ဖြစ်သည်။ ထို့အပြင် Multiple Linear Regression တွင် ရလဒ်ကို ခန့်မှန်းရာတွင် အသုံးပြုသည့် အချက်အလက်များမှာ တခုနှင့် တခု မှီခိုခြင်း နည်းနိုင်သရွေ့ နည်းရန် ဖြစ်သည်။ အကယ်၍ အထက်ပါ ယူဆချက်များနှင့် မကိုက်ညီပါက Linear Regression မှ ရလာသည့် ရလဒ်များမှာ တိကျမှု အားနည်းနေမည် ဖြစ်သည်။

Polynomial Regression

၁၅။ Polynomial Regression Analysis ဆိုသည်မှာ ရလဒ်နှင့် အချက်အလက်များ၏ ဆက်သွယ်ချက်ကို မျဉ်းကွေးဖြင့် ဖော်ပြခြင်းဖြစ်သည်။ သန်းခြာအားဖြင့် အောက်ပါအတိုင်း ဖော်ပြနိုင်သည်။

$$y = a_1 x^n + a_2 x^{\{n-1\}} + a_3 x^{\{n-2\}} + \ldots + a_n x + b \qquad (၂.၆)$$

၁၆။ အထက်ပါ ညီမျခြင်းကို ပုံဆွဲကြည့်မည်ဆိုပါက ပုံ (၂) တွင် ဖော်ပြထားသည့် မျဉ်းကွေးကို ရရှိပါသည်။ ညီမျခြင်း ၂.၆ ကို Polynomial equation ဟု ခေါ်ပြီး ညီမျခြင်း၏ degree သို့ order ဖြစ်သည့် n ၏ တန်ဖိုးပေါ်တွင် မူတည်၍ မျဉ်းကွေး၏ အနေအထားမှာ ပြောင်းလဲ သွားမည် ဖြစ်သည်။ n ၏ တန်ဖိုးကို ၁ ဟု သတ်မှတ်ပါက မျဉ်းဖြောင့်ကို ရရှိမည် ဖြစ်ပြီး n ၏ တန်ဖိုးမြင့်တက်လာသည်နှင့် အမျ ခန့်မှန်းရလဒ် (မျဉ်းကွေး)နှင့် မူလတန်ဖိုး (အပြာ အစက်များ) အကြား ကွာဟမှု နည်းပါးလာမည် ဖြစ်သည်။ သို့သော် အခြားတဖက်ကမူ ရှာဖွေရမည့် ကိန်းသေ (parameter) အရေအတွက်များပြားလာမည် ဖြစ်သည်။



ပုံ (၂) Illustration of Polynomial Regression

Model Implementation

၁၇။ Regression model ကို Python Library များ အသုံးပြု၍ အလွယ်တကူ တည်ဆောက်နိုင်ပါသည်။ Machine learning model တခု မတည်ဆောက်မီ ဦးစွာ အရေးကြီးသည့် အချက်မှာ ဒေတာများကို ပြင်ဆင်ရန် ဖြစ်ပါသည်။ မူလ အချက်အလက်များ မှားယွင်းနေပါက ရလာသည့် ရလဒ်သည်လည်း မည်သို့မှ မတိကျနိုင်ပါ။ ထို့အပြင် ရလဒ်ကို ခန့်မှန်းရာတွင် အသုံးပြုမည့် အချက်အလက်များကို ရွေးချယ်ရာတွင်လည်း မှန်ကန်သည့် အချက်အလက်များ ဖြစ်ရန် လိုအပ်သည်။ ဥပမာ ကုမ္ပဏီ တခုသည် သူ၏ ကုန်ပစ္စည်းများကို ရုပ်မြင်သံကြား၊ ရေဒီယိုနှင့် သတင်းစာများတွင် ကြော်ငြာလေ့သည် ဆိုကြပါစို့။ အဆိုပါ ကြော်ငြာများအတွက် ကုန်ကျသည့် ကုန်ကျစရိတ်နှင့် ကုမ္ပဏီ၏ ရောင်းအား ဆက်သွယ်မှုကို သိရှိနိုင်ရန် နေ့စဉ် (သို့) အပတ်စဉ် အတွက် အစီအစဉ်တခုချင်းစီအတွက် ကုန်ကျစရိတ်နဲ့ ထိုကာလ အတွင်း ရောင်းရသည့် ရောင်းအားများပါဝင်သည့် data-set တခုကို ဦးစွာတည်ဆောက် ရမည် ဖြစ်သည်။ ထိုသို့ ကောက်ယူရာတွင်လည်း ၁ ပတ်စာ ၊ တလစာ ကောက်ယူရုံဖြင့် မလုံလောက်ပဲ အနည်းဆုံး တနှစ်စာ ပါဝင်သည့် ဒေတာ အချက်အလက်များ ကောက်ယူရန်လိုအပ်ပါသည်။

၁၈။ supervised regression method သည် ကြိုတင် စုဆောင်းထားသည့် အချက်အလက်များကို အသုံးပြု၍ ရလဒ်နှင့် အချက်အလက်ကြား ဆက်သွယ်မှုကို ဖော်ပြနိုင်မည့် ကိန်းသေ (parameter)များကို ရှာဖွေခြင်း ဖြစ်သည်။ machine learning model တည်ဆောက်ရာတွင် စုဆောင်းထားသည့် အချက်အလက်အားလုံးကို အသုံးပြုခဲ့မည် ဆိုပါက model ၏ လုပ်ဆောင်ချက်ကို ပြန်လည် ဆန်းစစ်ရန် ခက်ခဲပါလိမ့်မည်။ သို့ဖြစ်ရာ ကြိုတင် စုဆောင်းထားသည့် အချက်အလက်များအနက်မှ တချို့တဝက်ကို model တည်ဆောက်ရာတွင် အသုံးမပြုပဲ ချန်လှပ်ထားရန် လိုအပ်ပါသည်။ ဥပမာ အထက်ပါ ကုမ္ပဏီမှ ရက် ၃၀၀ စာ ဒေတာ ကောက်ယူထားသည် ဆိုပါက ရက် ၂၀၀ စာ ဒေတာကို machine learning model တည်ဆောက်ရာတွင် အသုံးပြု၍ ကျန် ရက် ၁၀၀ စာကို model ၏ လုပ်ဆောင်ချက်ကို ပြန်လည် ဆန်းစစ်ရန်အတွက် အသုံးပြုနိုင်ပါသည်။ အဆိုပါ ရက် ၂၀၀ စာ ဒေတာကို Training data set ဟု ခေါ်ဆို၍ ကျန် ရက် ၁၀၀ စာ ဒေတာကို Testing data set ဟု ခေါ်ဝေါ် သုံးစွဲသွားမည် ဖြစ်သည်။

၁၉။ အချက်အလက်များ အဆင်သင့်ဖြစ်ပြီးဆိုပါက Training data set ကို အသုံးပြု၍ machine learning model စတင် တည်ဆောက်နိုင်မည် ဖြစ်ပါသည်။ ထို့နောက် Testing data set ကို အသုံးပြု၍ machine learning model ၏ လုပ်ဆောင်ချက်ကို ပြန်လည် ဆန်းစစ်ရပါမည်။ သို့မှသာ machine learning ကို လက်တွေ့ အသုံးချရာတွင် ရရှိနိုင်မည့် ရလဒ်ကို ကြိုတင် ခန့်မှန်းနိုင်မည် ဖြစ်သည်။