One hot encoding زمانی استفاده میشود که داده های دسته ای یا کتوگوریکال را به داده باینری تبدیل کنیم، از معایب ان زمانی که دسته ها زیاد باشد باعث افزایش فیچر های مورد نظر و پیچیدگی مدل میشود

با استفاده از تبدیل log and exponential transformation میتوان داده های با توزیع log normal را به normal تبدیل کرد.

روش bootstrapping از روشهای نمونه گیری مجدد است و روش کارآمدی برای محاسبه میزان دقت و خطای استاندارد (standard error) متغیر تخمین زده شده است . در روش bootstrappingبه صورت تصادفی n عضو با جایگذاری انتخاب میکنیم
برای مثال فرض کنید مقدار معینی پول داریم که م یخواهیم بخش ی α درصد آن را در یک پروژه و $1-\alpha$درصد آن را در یک پروژه دیگر سرمای هگذاری کنیم .اگر میزان سود پروژه اول X و میزان سود پروژه دوم Y باشد ، به دنبال یافتنα ای هستیم که به ازای آن، میزان ریسکمان یعنی $Var(\alpha X+(1-\alpha)Y)$ کمینه شود

در روش cross validation هر بار یک زیرمجموعه از داد ههای آموزشی را کنار م یگذاریم، مدل را با استفاده از بقیه داد ههها آموزش میدهیم و از داد ههای کنار گذاشته شده به عنوان داد ههای تست استفاده میکنیم.
در 2-fold cross-validation داده ها را بصورت رندوم به 2 بخش هم اندازه تقسیم می کنیم .یک بخش به عنوان داد هی ارزیابی و دیگری بعنوان داد هی آموزشی استفاده می شود و بار دیگر همین کار را با بخش دیگر انجام می دهیم در نتیجه دو تا معیار خطای تست خواهیم داشت که می توانیم از آن ها میانگین بگیریم و بعنوا ن خطای تست مدل گزارش دهیم.مزیت :هر بخش یکبار بعنوان داد هی ارزیابی استفاده شده است واز همه ی مشاهدات استفاده کرده ایم.
در 5*2 fold cross-validation ، ۵ بار 2-fold cross-validation را اجرا می کنیم .می توان خطایتست در این مورد با میانگین گرفتن از ۵ نتیج هی 2 fold cross-validation بدست آورد. این روش که در واقع تکرار روش k-fold cv است باعث می شود تا تخمین بهتری از خطای تست بدست آید

براساس means-k ومفهوم متغییر inertiaموجود دران که برابر مجموع مربعات فاصله ها از نزدیک ترین مرکز خوشه است. وقتی نمودار مقادیر k بر حسب inertia را رسم شود از مقدار k مشخص به

بعد نمودار به صورت خطی کم می شود. ولی شرایط قبل نمایی است. به این نقطه elbowگفته میشود. و مقدار بهینه برای k این نقطه است. در نمودار زیر نیز چنین الگویی وجود دارد. در نمودار از مرتبه ۳ تا ۵ مقدار خطی است و بنابراین بهترین مدل برای پیچیدگی پیدا میشود.
رفتار تغییر بایاس همیشه به صورت خطی نیست اما با مشاهده واریانس و افزایش واریانس نسبت به کاهش بایاس می توان نقطه بهینه برای انتخاب پیچیدگی مدل را پیدا کرد پس نظر به طور کلی نمیتوان گفت.