Makine öğrenmesi uygulamalarında temel amaç eldeki veriler üzerinden örüntüler elde etmek ve yeni veriler için bu örüntüler üzerinden doğru tahminler yapabilmektir. Bu tahminleri yapabilmek için makine öğrenmesi uygulamaları sonucunda bir model elde ederiz. Peki model nedir ?

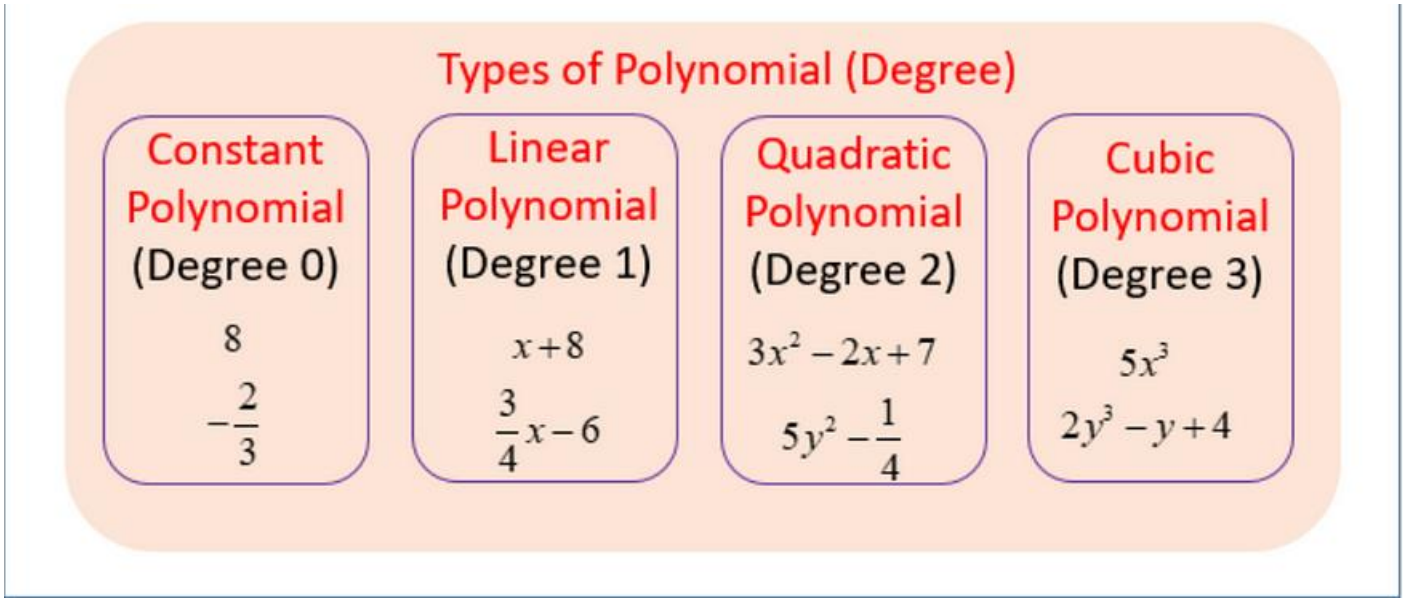
Model *girdilerin çıktılara eşlenmesi için kullanılan bir sistemdir. Örneğin amacımız ev fiyatlarını tahmin etmek olsun . Bunun için evin metrekare bilgisini girdi olarak kullanan bir model oluştururuz ve çıktı olarak da evin fiyatını elde ederiz.*

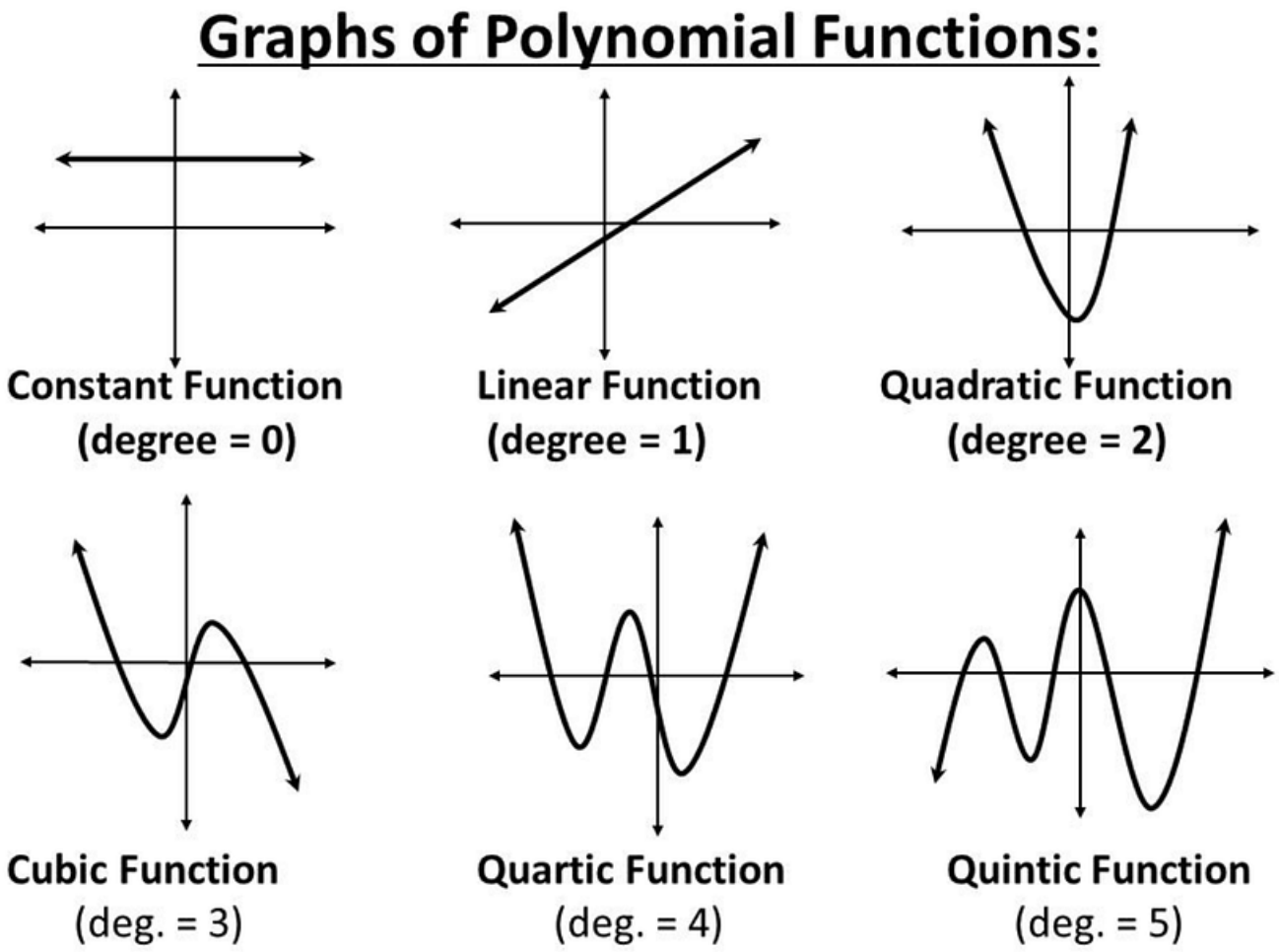
Bir model bir problem hakkında teori üretir. Buradaki teori evin metrekare bilgisi ile fiyatı arasında bir ilişki olmasıdır. Eğitim veri çalışmaları sonucunda bu ilişkiyi öğrenmiş olan bir model oluşturmuş oluruz.

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_n x^n + \varepsilon.$$

Burada **y** tahminlenen değer ve **x'ler** ise modelin değişkenleridir. Beta terimleri ise model parametreleridir ve bu parametre değerleri eğitim seti üzerinde yapılan çalışmalar sonucunda öğrenilmektedir. epsilon ise modelin hata değeridir. Model Beta değerlerini öğrendikten sonra istediğimiz x değişkenlerini formüle koyarak y değerini bulabiliriz.

Modelimizdeki değişken sayısına bağlı olarak (degree) farklı grafikler elde ederiz.





Makine öğreniminde gelecek veriler hakkında tahmin yapabilmek için verilerimizi eğitim verileri (train) ve test verileri olmak üzere iki alt kümeye ayırıyoruz. Modelimizi eğitim verilerinden elde edilen örüntülere göre oluşturuyoruz. Bu işlem sonucunda iki şeyden biri olabilir; modelimiz aşırı öğrenebilir veya eksik öğrenebilir. Bu durumda modelimiz yeterli öngörüde bulunamayacak ve tahminlerimizde hata oranı yüksek olacaktır.

Overfitting

Eğer modelimiz, eğitim için kullandığımız veri setimiz üzerinde gereğinden fazla çalışıp ezber yapmaya başlamışsa ya da eğitim setimiz tek düze ise **overfitting** olma riski büyük demektir. Eğitim setinde yüksek bir skor aldığımız bu modele, test verimizi gösterdiğimizde muhtemelen çok düşük bir skor elde edeceğiz. Çünkü model eğitim setindeki durumları ezberlemiştir ve test veri setinde bu durumları aramaktadır. En ufak bir değişiklikte ezberlenen durumlar bulunamayacağı için test veri setinde çok kötü tahmin skorları elde edebilirsiniz. Overfitting problemi olan modellerde yüksek varyans, düşük bias durumu görülmektedir.

Bu genellikle model çok karmaşık olduğunda (yani gözlem sayısına kıyasla çok fazla özellik / değişken varsa) gerçekleşir. Bu model eğitim verilerinde çok yüksek tahmin doğruluğuna sahip olacaktır, ancak eğitimsiz veya yeni verilerde muhtemelen çok doğru tahminleme yapamayacaktır. Bu sorun modelin genelleştirme yapamamasından kaynaklanmaktadır. Bu tip modeller verilerdeki değişkenler arasındaki gerçek ilişkiler yerine eğitim verilerindeki “gürültüyü” öğrenir veya açıklar.

Overfitting problemi aşağıdaki yöntemler uygulanarak çözülebilmektedir;

- **Öz nitelik sayısını azaltmak:** Birbirleriyle yüksek korelasyonlu olan kolonlar silinebilir ya da faktör analizi gibi yöntemlerle bu değişkenlerden tek bir değişken oluşturulabilir.
- **Daha fazla veri eklemek :** Eğer eğitim seti tek düze ise daha fazla veri ekleyerek veri çeşitliliği arttırılır.
- **Regularization (Düzenleme) :** Düzenleme, modelin karmaşıklığını azaltmak için bir kullanılan tekniktir. Bunu kayıp fonksiyonunu cezalandırarak yapar. Yani modelde ağırlığı yüksek olan değişkenlerin ağırlığını azaltarak bu değişkenlerin etki oranını azaltır. Bu yöntem, aşırı öğrenme probleminin çözülmesine yardımcı olur. Kayıp fonksiyonu, gerçek değer ile öngörülen değer arasındaki farkın karelerinin toplamıdır. Değişkenlerin ağırlığını azaltmak için regularization değerini arttırmak gerekmektedir. En popüler Regularization metotları **Lasso** ve **Ridge** tekinkleridir.

Underfitting

Aşırı öğrenmenin aksine, bir model yetersiz öğrenmeye sahipse, modelin eğitim verilerine uymadığı ve bu nedenle verilerdeki trendleri kaçırdığı anlamına gelir. Ayrıca modelin yeni veriler için genelleştirilemediği anlamına da gelir. Tahmin ettiğiniz gibi bu problem genellikle çok basit bir modelin sonucudur (yetersiz tahminleyici bağımsız değişken eksikliği).

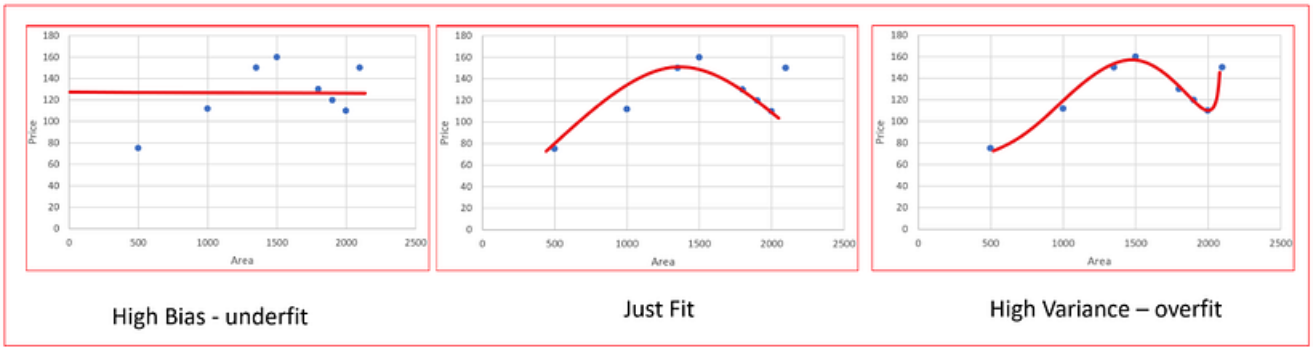
Underfitting sorunu olan modellerde hem eğitim hem de test veri setinde hata oranı yüksektir. Düşük varyans ve yüksek bias'a sahiptir. Bu modeller eğitim verilerini çok yakından takip etmek yerine, eğitim verilerinden alınan dersleri yok sayar ve girdiler ile çıktılar arasındaki temel ilişkiyi öğrenemez.

Underfit'in overfit kadar yaygın olmadığını belirtmek gerekir. Yine de, veri analizinde her iki problemden de kaçınmak gereklidir.

Varyans-Bias Çelişkisi

Varyans, model eğitim veri setinde iyi performans gösterdiğinde, ancak bir test veri kümesi veya doğrulama veri kümesi gibi, eğitilmemiş bir veri kümesinde iyi performans göstermediğinde ortaya çıkar. Varyans, gerçek değerden tahmin edilen değerin ne kadar dağınık olduğunu söyler.

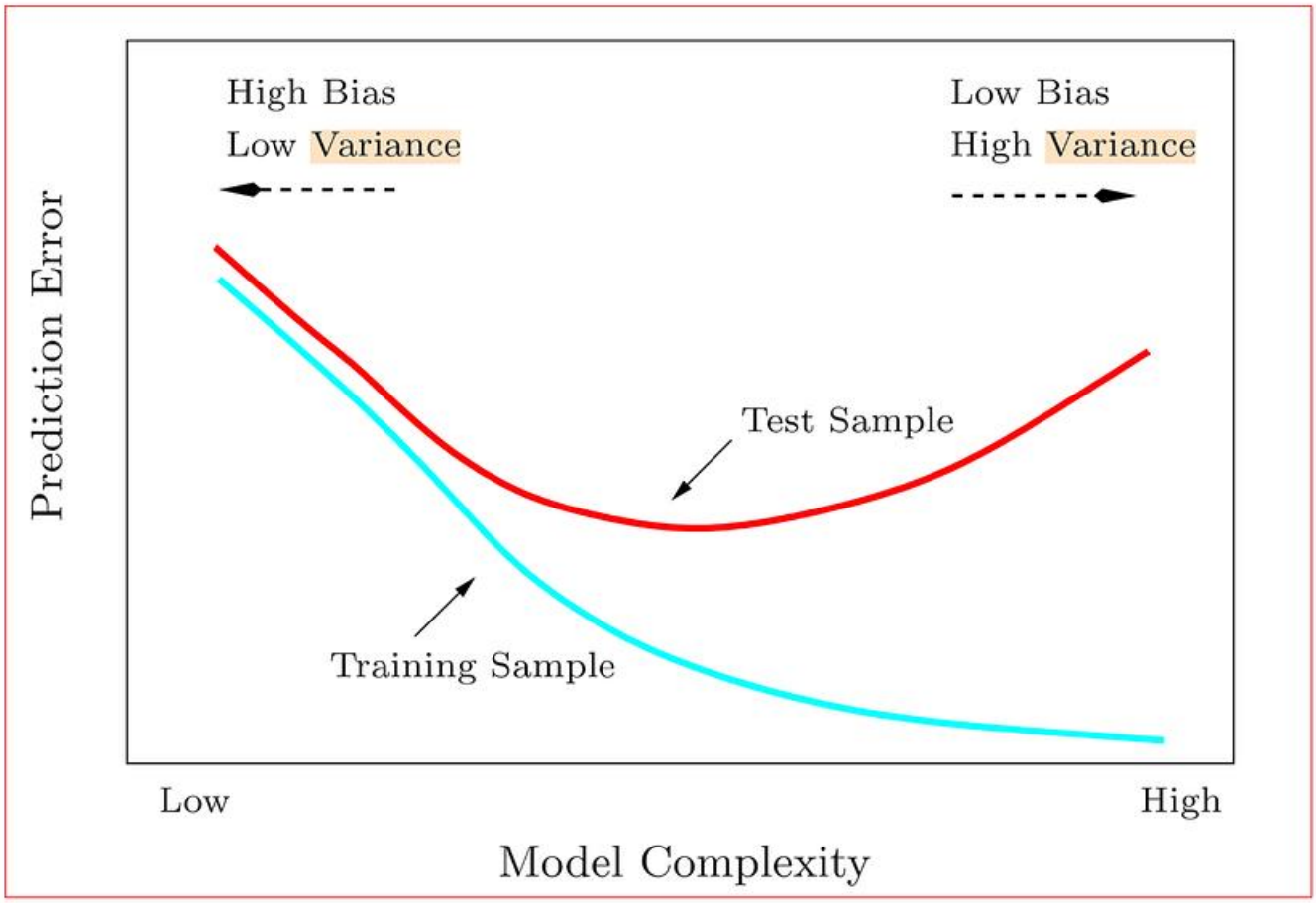
Bias, gerçek değerlerden tahmin edilen değerlerin ne kadar uzak olduğudur. Tahmin edilen değerler gerçek değerlerden uzaksa, bias yüksektir.



High bias , high variance and just fit

Yukarıdaki grafiklere bakarsak, yüksek bias'a sahip bir modelin çok basit olduğu görülmektedir. Yüksek varyansa sahip bir model, veri noktalarının çoğuna uymaya çalışır ve bu da modeli karmaşık yapar ve modellenmesini zorlaştırır.

Aşağıdaki grafikten görüldüğü gibi model karmaşıklığı arttıkça eğitim seti üzerinde hatalı tahmin oranı azaltmakta ancak test veri seti üzerinde tahmin hatası artmaktadır.



Source: Elements of Statistical Learning by Trevor Hastie, Robert Tibshirani and Jerome Friedman

- Yüksek Bias Düşük Varyans: Modeller tutarlıdır, ancak ortalama hata oranı yüksektir.
- Yüksek Bias Yüksek Varyans : Modeller hem hatalı hem de tutarsızdır .
- Düşük Bias Düşük Varyans: Modeller ortalama olarak doğru ve tutarlıdır. Modellerimizde bu sonucu elde etmek için çabalamaktayız.
- Düşük Bias Yüksek Varyans: Modeller bir dereceye kadar doğrudur ancak ortalamada tutarsızdır. Veri setinde ufak bir değişiklik yapıldığında büyük hata oranına neden olmaktadır.

Yüksek bias veya yüksek varyansa sahip olduğumuzu bulmanın yolu nedir?

Eğer model yüksek bias'a sahipse aşağıdaki sonuçlarla karşılaşmamız kaçınılmazdır;

- Modelin eğitim setinin hata oranı yüksektir.
- Test / doğrulama veri seti hata oranı eğitim seti ile benzer oranda yüksektir.

Eğer model yüksek varyans'a sahipse aşağıdaki sonuçlarla karşılaşmamız kaçınılmazdır;

- Modelin eğitim setinin hata oranı düşüktür.
- Modelin test/doğrulama veri setinin hata oranı yüksektir.

Yüksek bias problemini çözmek için aşağıdaki yöntemleri uygulayabiliriz.

- **Daha fazla veri eklemek :** Daha fazla veri ekleyerek veri çeşitliliğini arttırmak gereklidir.
- **Daha fazla değişken eklemek :** Model karmaşıklığının artmasını sağlamaktadır.
- **Regularization (düzenleme) :** Değişkenlerin ağırlığını arttırmak için regularization değerini azaltın.