Week 10: Evaluating and Interpreting Linear Regression, Machine Learning

DSUA111: Data Science for Everyone, NYU, Fall 2020

TA Jeff, `jpj251@nyu.edu`

- This slideshow: https://jjacobs.me/dsua111-sections/week-10 (https://jjacobs.me/dsua111-sections/week-10)
- All materials: https://github.com/jpowerj/dsua111-sections (https://github.com/jpowerj/dsua111-sections)

Outline

I. Evaluating and Interpreting Linear Regression

1. Coefficient Interpretation
2. p-value Interpretation
3. Evaluating Results

II. Machine Learning (aka Fancy Regressions)

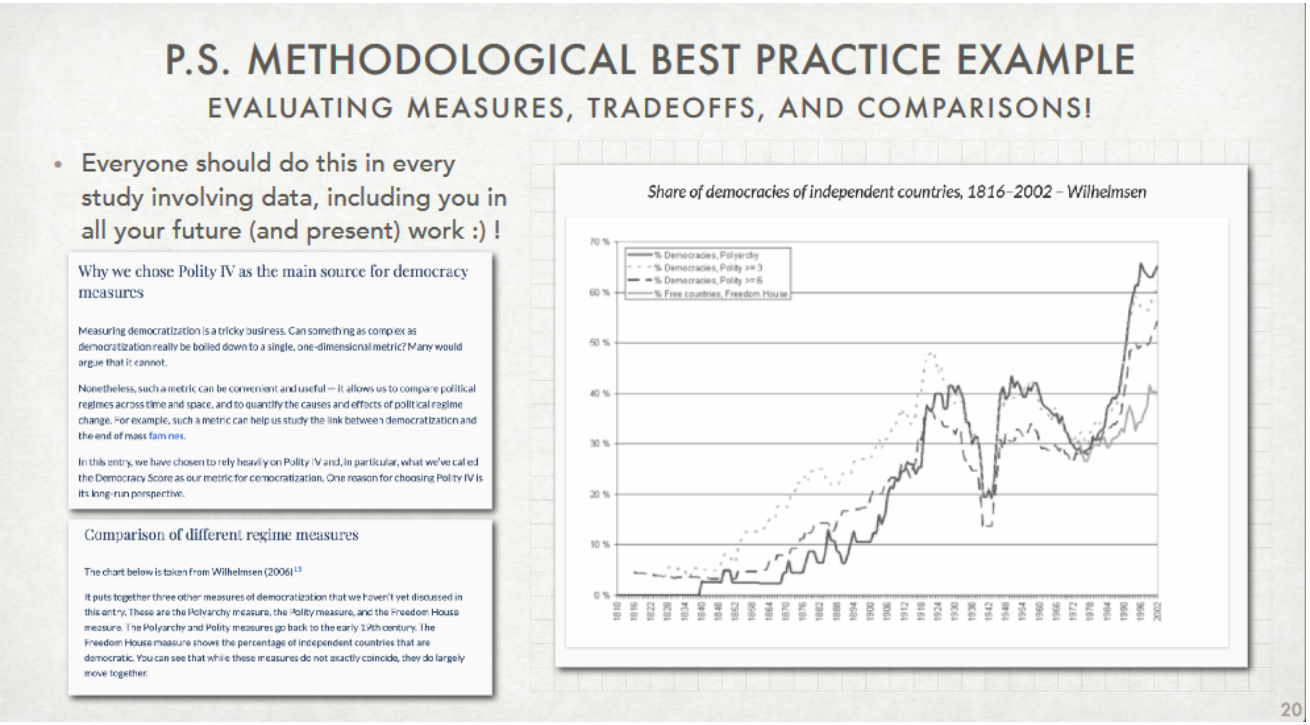
1. Statistics vs. Machine Learning?
2. ML Approach to Regression
3. Evaluating Machine Learning Models

Part I: Evaluating and Interpreting Linear Regression

Substantive vs. *Significant* Effects

- ***Substantive***: Is it interesting for society?
- ***Significant***: How much do we believe it?
- Example: A government carries out two studies:
    - One finds that a means-tested welfare program decreases poverty by 0.00001%, significant at the 1% level of confidence.
    - The other finds that a universal basic income program decreases poverty by 10%, significant at the 10% level of confidence (but not 5% or 1%)
- Which is more substantive? Which is more significant?
- October 16, 2020: "JFI, Mayor Aja Brown, and Fund for Guaranteed Income announce largest city-based guaranteed income initiative" (https://www.jainfamilyinstitute.org/news/jfi-mayor-aja-brown-and-fund-for-guaranteed-income-announce-largest-city-based-guaranteed-income-initiative/)

Justify Your Measures!



P.S. METHODOLOGICAL BEST PRACTICE EXAMPLE

EVALUATING MEASURES, TRADEOFFS, AND COMPARISONS!

- Everyone should do this in every study involving data, including you in all your future (and present) work :) !

Why we chose Polity IV as the main source for democracy measures

Measuring democratization is a tricky business. Can something as complex as democratization really be boiled down to a single, one-dimensional metric? Many would argue that it cannot.

Nonetheless, such a metric can be convenient and useful — it allows us to compare political regimes across time and space, and to quantify the causes and effects of political regime change. For example, such a metric can help us study the link between democratization and the end of mass famines.

In this entry, we have chosen to rely heavily on Polity IV and, in particular, what we've called the Democracy Score as our metric for democratization. One reason for choosing Polity IV is its long-run perspective.

Comparison of different regime measures

The chart below is taken from Wilhelmsen (2006)[15]

It puts together three other measures of democratization that we haven't yet discussed in this entry. These are the Polyarchy measure, the Polity measure, and the Freedom House measure. The Polyarchy and Polity measures go back to the early 19th century. The Freedom House measure shows the percentage of independent countries that are democratic. You can see that while these measures do not exactly coincide, they do largely move together.

Share of democracies of independent countries, 1816–2002 – Wilhelmsen

p-values

[In theory,] **"the probability of getting the coefficient we got if the null hypothesis is correct"**

- In regression analyses, null hypothesis is almost always "coefficient of interest is 0"
- Hence, in regression the p-value is "the probability we'd get the *observed coefficient* (the coefficient in the regression results table) if the *true coefficient* is 0"
- $\implies$ the smaller the p-value is, the more we "believe" that the effect we found is real

In practice, too many "moving parts" in experiments (and ESPECIALLY in observational studies) for this to be that meaningful

- Bayesian statistical approach used more and more since ~the 80s, since computers allow us to compute more subtle measures/incorporate more details and prior knowledge than the (frequentist) p-value
- e.g., if 1 million experiments up til now found the gravitational constant to be 6.67408×10^{-11}, but we do an experiment which results in an estimate of 3, Bayesian approach allows us to view this result more skeptically given *prior* information/data
- Gelman, Andrew. "P Values and Statistical Practice". *Epidemiology* 24(1): 69-72. 2013 (http://www.stat.columbia.edu/~gelman/research/published/pvalues3.pdf)

OLS Diagnostics

In this class we focus on:

- R-squared
- Adj. R-squared
- Prob(F-statistic)
- Confidence interval

R-Squared

- In theory, **"The proportion of variance in the dependent variable 'explained by' the independent variable(s)"**
- In practice, "explained by" is a loaded term, and honestly I would just think of it like a regression-specific correlation measure
- "Adjusted" R-Squared just applies a penalty for each additional variable you include in your model (since you can always make non-adjusted R-squared increase by adding more and more variables)

F-Statistic

- The p-value is specific to *one* coefficient, on one independent variable ("how significant is this coefficient/this independent variable's effect on the dependent variable")
- The F-statistic can be thought of like a p-value for your *whole* model, for *all* coefficients/independent variables. Similarly to the p-value, the F-statistic value is the probability of getting your model if *all* coefficients were 0.

Confidence Intervals

- "Confidence intervals tell us: **If we conduct our study many, many times, 95% of the time our confidence interval will capture the true coefficient**"
- "Important: Confidence intervals **do not** say that we are **'95% confident our coefficient is within this range'**"
- (Jeff's 2 cents: more trouble than they're worth... see Gelman article above)

Multiple Regression

- You can have more than one independent variable!
- $Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \ldots + \beta_N x_{N,i}$
    - $x_{1,i}$: 1st independent variable (its value for person i)
    - $x_{2,i}$: 2nd independent variable (its value for person i)
    - $\ldots$
    - $x_{N,i}$: Nth independent variable (its value for person i)
- Now, the coefficient interpretation changes a bit:
    - An increase of 1 unit in x_1 is associated with an increase of β_1 units in Y, **when all other variables are held constant**

Part II: Machine Learning (aka Fancy Regressions)

Statistics vs. Machine Learning

Biggest difference has to do with the respective goals of the two approaches:

- **Statistics**: Inference (Why is something happening? What is the relationship between variables of interest? What causes economic growth?)
- **Machine Learning**: Prediction (What is going to happen next? Will people who do X also do Y? What will be the level of the stock market one month from today?

(from Lecture 18.1, Slide 4)

The Difference in Practice

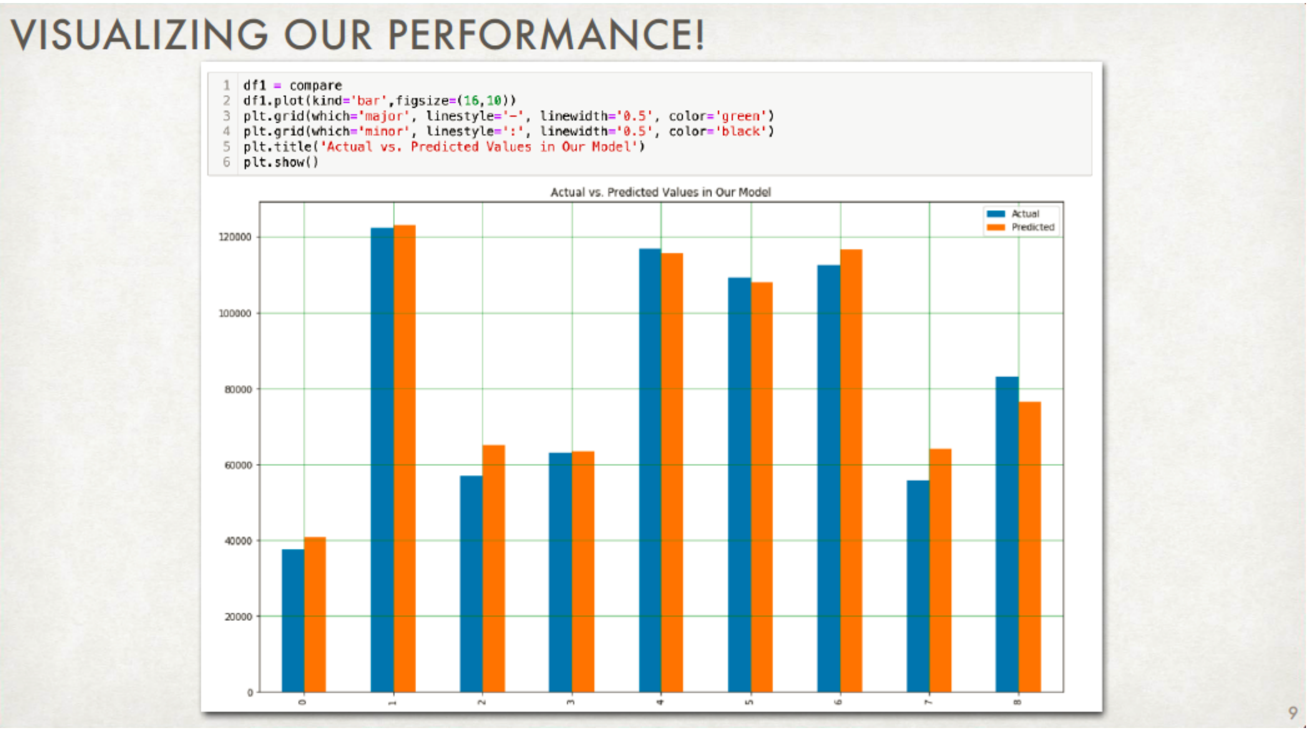
- You've already seen "statistical" regression, where you toss all the data into `statsmodels` and observe the result
- In Machine Learning, it's a bit more involved:
    1. Split the data into training and testing sets (typically 80% of the data for training, 20% for testing)
    2. **Train** the model on the training set
    3. **Evaluate** the model on the testing set
    4. Make **predictions** (but perhaps don't learn as much about the underlying relationships between variables)
- [Note: SUPER important that the ML algorithm **never** sees the test data]

(from Lecture 18.1, Slide 10)

`scikit-learn`

1. `import sklearn`
2. Load data
3. Split into training (80%) and test (20%) sets
4. **Train** the model
    - (What's going on "under the hood" here? Actually not that scary and worth knowing!)
5. Ask the model to predict **dependent variable** values given **independent variable values** in the **test set**
6. Compute **accuracy** by comparing these **predicted** dependent variable values with the **actual** dependent variable values in the test set (important: the ML model never ever looks at these!)

Computing Accuracy

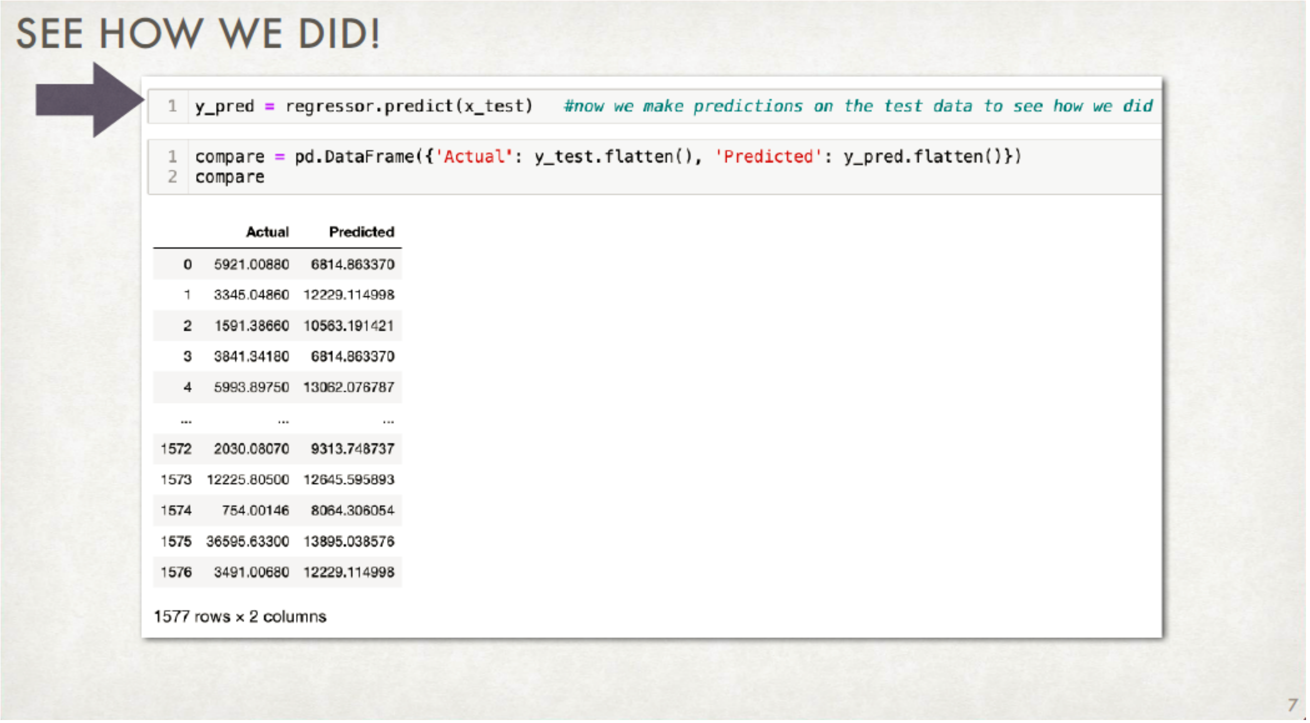


VISUALIZING OUR PERFORMANCE!

```
df1 = compare
df1.plot(kind='bar',figsize=(16,10))
plt.grid(which='major', linestyle='-', linewidth='0.5', color='green')
plt.grid(which='minor', linestyle=':', linewidth='0.5', color='black')
plt.title('Actual vs. Predicted Values in Our Model')
plt.show()
```

(Lecture 18.2, Slide 9)

Surprisingly Helpful: Predicted vs. Actual Values Table



SEE HOW WE DID!

```python
y_pred = regressor.predict(x_test)    #now we make predictions on the test data to see how we did
```

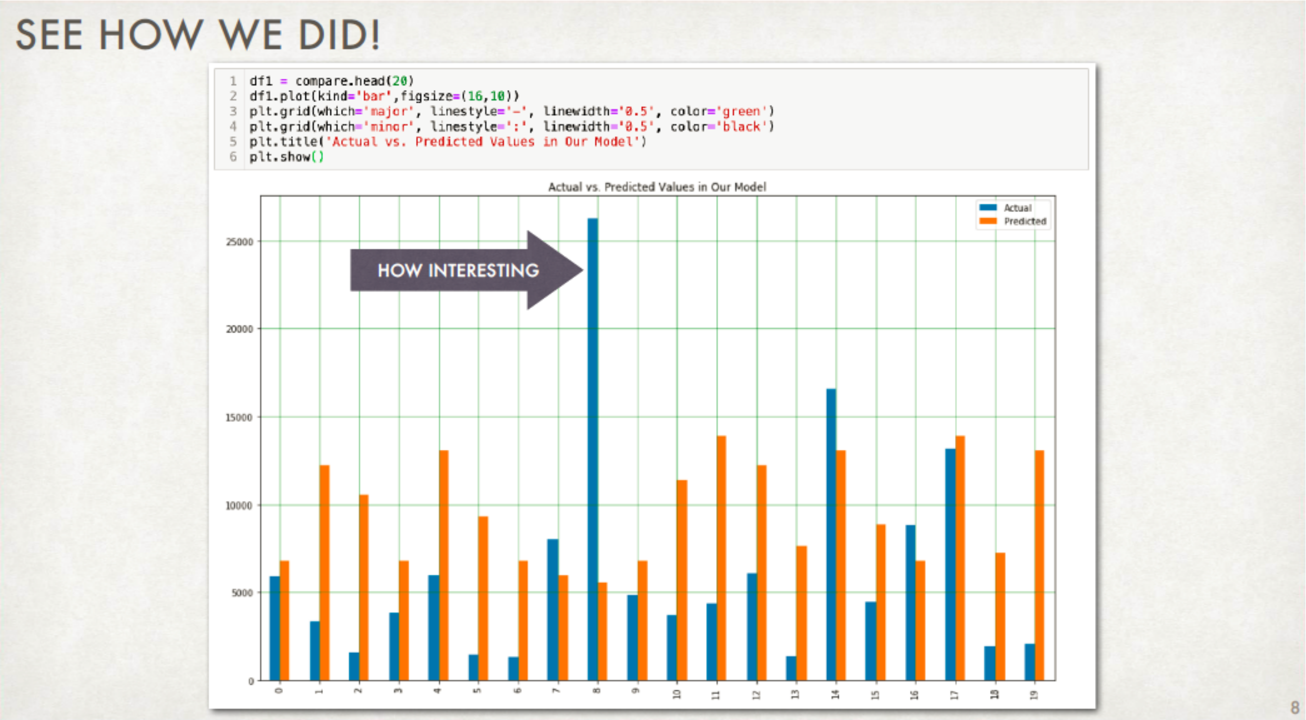
```python
compare = pd.DataFrame({'Actual': y_test.flatten(), 'Predicted': y_pred.flatten()})
compare
```

	Actual	Predicted
0	5921.00880	6814.863370
1	3345.04860	12229.114998
2	1591.38660	10563.191421
3	3841.34180	6814.863370
4	5993.89750	13062.076787
...	...	...
1572	2030.08070	9313.748737
1573	12225.80500	12645.595893
1574	754.00146	8064.306054
1575	36595.63300	13895.038576
1576	3491.00680	12229.114998

1577 rows × 2 columns

(Lecture 18.3, Slide 7)

Less-Surprisingly Helpful: Visualizing Accuracy

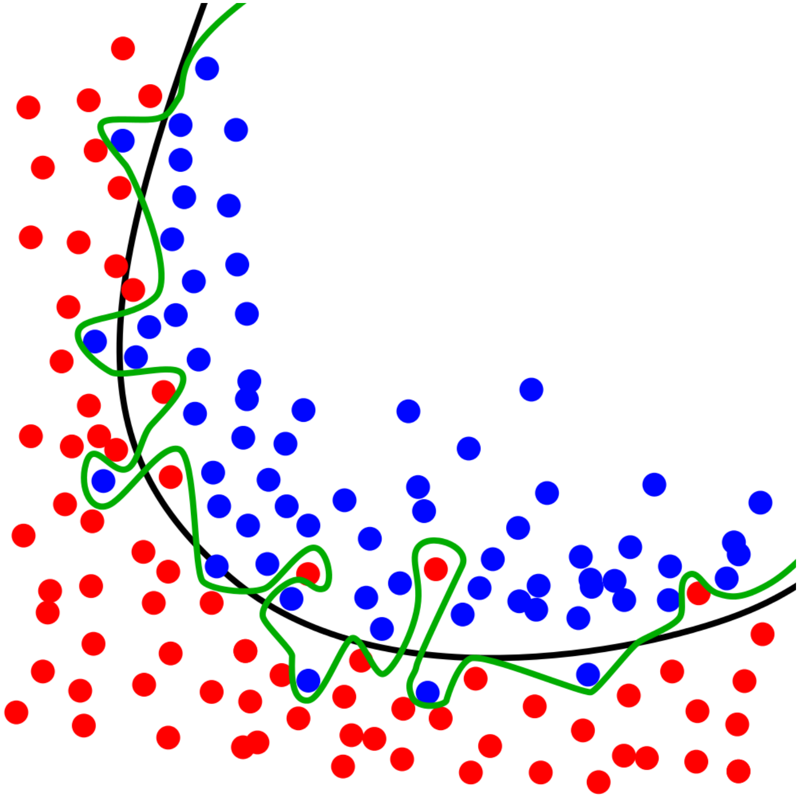


SEE HOW WE DID!

```python
df1 = compare.head(20)
df1.plot(kind='bar',figsize=(16,10))
plt.grid(which='major', linestyle='-', linewidth='0.5', color='green')
plt.grid(which='minor', linestyle=':', linewidth='0.5', color='black')
plt.title('Actual vs. Predicted Values in Our Model')
plt.show()
```

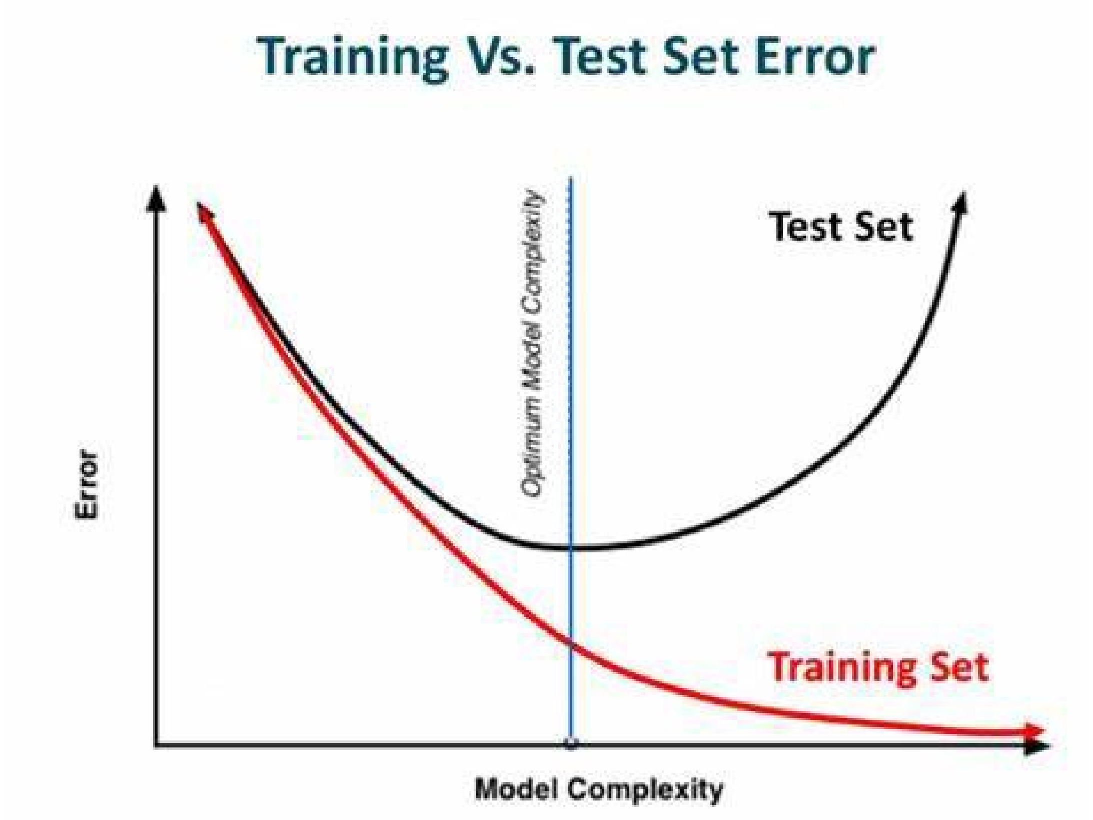
(Lecture 18.3, Slide 8)

OVERFITTING: Why you should NOT be maximizing accuracy



(from https://en.wikipedia.org/wiki/Overfitting (https://en.wikipedia.org/wiki/Overfitting))

How to Know When You’re Overfitting



(from https://medium.com/ai%C2%B3-theory-practice-business/dropout-early-stopping-188a23beb2f9 (https://medium.com/ai%C2%B3-theory-practice-business/dropout-early-stopping-188a23beb2f9))