Berkeley Engineering | BerkeleyHaas

PROFESSIONAL CERTIFICATE IN MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE

Office Hour #8 with Matilde D'Amelio

May 5, 2022 at 9 pm UTC

FEATURE ENGINEERING

Feature Engineering is the process of taking certain variables (features) from our dataset and transforming them in a predictive model. Essentially, we will be trying to manipulate single variables and combinations of variables in order to *engineer* new features. By creating these new features, we are increasing the likelihood that one of the new variables has more predictive power over our outcome variable than the original, un-transformed variables.

FEATURE ENGINEERING

Goal: predict carseat sales with a good R-squared

Step 1: we look at the covariance to eliminate some variables + calculate R-squared

	Urban	US	Good	Medium	CompPrice	Income	Advertising	Population	Price	Age	Education
Urban	0.207298	0.007612	-0.004045	-0.003321	0.040568	0.020323	0.009927	-0.016719	0.044636	0.024700	-0.019609
US	0.007612	0.221375	0.018696	-0.009977	0.001238	0.038094	0.301661	0.041366	-0.007492	0.000218	-0.060227
Good	-0.004045	0.018696	0.168840	-0.111767	0.027146	-0.007704	0.035925	0.020407	0.010887	-0.002451	0.001401
Medium	-0.003321	-0.009977	-0.111767	0.250513	0.015909	-0.027210	-0.015691	-0.043513	0.006767	0.043082	0.012606
CompPrice	0.040568	0.001238	0.027146	0.015909	0.886988	-0.095025	-0.055005	-0.119792	0.478431	-0.125330	-0.014849
Income	0.020323	0.038094	-0.007704	-0.027210	-0.095025	1.021401	0.073764	0.052228	-0.008790	-0.049898	-0.047055
Advertising	0.009927	0.301661	0.035925	-0.015691	-0.055005	0.073764	0.930377	0.249023	0.002019	-0.014507	-0.015874
Population	-0.016719	0.041366	0.020407	-0.043513	-0.119792	0.052228	0.249023	1.005038	-0.033357	-0.048807	-0.156006
Price	0.044636	-0.007492	0.010887	0.006767	0.478431	-0.008790	0.002019	-0.033357	0.863050	-0.104037	0.001088
Age	0.024700	0.000218	-0.002451	0.043082	-0.125330	-0.049898	-0.014507	-0.048807	-0.104037	1.003598	0.001754
Education	-0.019609	-0.060227	0.001401	0.012606	-0.014849	-0.047055	-0.015874	-0.156006	0.001088	0.001754	1.019096

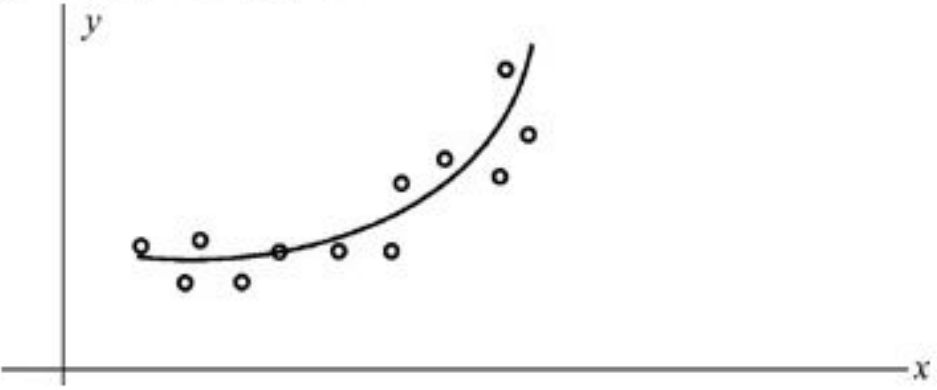
Step 2: Feature Engineering - try to create a wide variety of interactions between multiple variables in order to create new variables (we will create as many bivariate combinations)

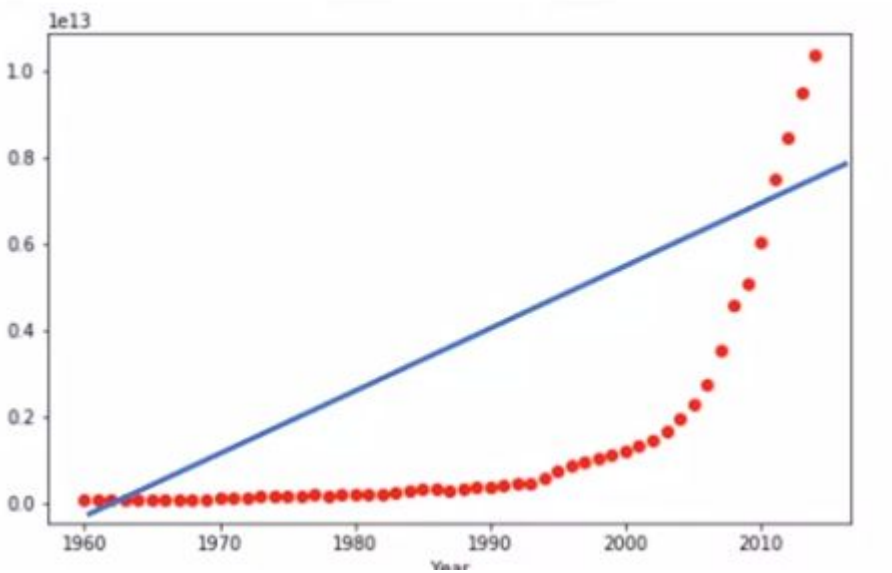
Step 3: we create a iterative linear regression that test all new features to identify the most performing features

Step4: we run a new regression with the most performing features and we calculate the R-squared

Parabolic Model Fitting & Non-Linear Feature

$$y = ax^2 + bx + c$$





Model's goal: Prediction or Inference

Inference

It refers to the act of reaching a conclusion that has been evaluated based on existing data, facts, and evidence. It involves building a model that describes the relationship between the variables and the outcome of an event or occurrence using statistical data.

There is a fair degree of certainty as the evaluation which has been conducted is factual.

Prediction

It refers to a conclusive statement made about a future event or occurrence.

There is a lower degree of certainty as the future is unknown.

Group Discussion

Business Examples of when you need to predict and when you need to interfere and you need to predict

(10 mins)

Overfitting

Overfitting refers to a model that models the training data too well.

Overfitting happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data. This means that the noise or random fluctuations in the training data is picked up and learned as concepts by the model. The problem is that these concepts do not apply to new data and negatively impact the models ability to generalize.

Overfitting is more likely with nonparametric and nonlinear models that have more flexibility when learning a target function. As such, many nonparametric machine learning algorithms also include parameters or techniques to limit and constrain how much detail the model learns.

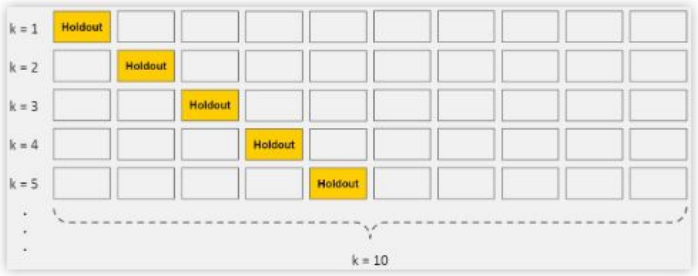
Prevent Overfitting

Get More Training Data

Cross-Validation & HoldOut: Use your initial training data to generate multiple mini train-test splits. Use these splits to tune your mode

Augmentation: If you can't get more data, you can try augmentation to add variation in your data. Augmentation means artificially modifying your existing data by means of transforms that resemble the variation you might expect in the real data.

Feature Selection: it is a process by which you automatically search for the best subset of attributes in your dataset.



QUESTIONS?