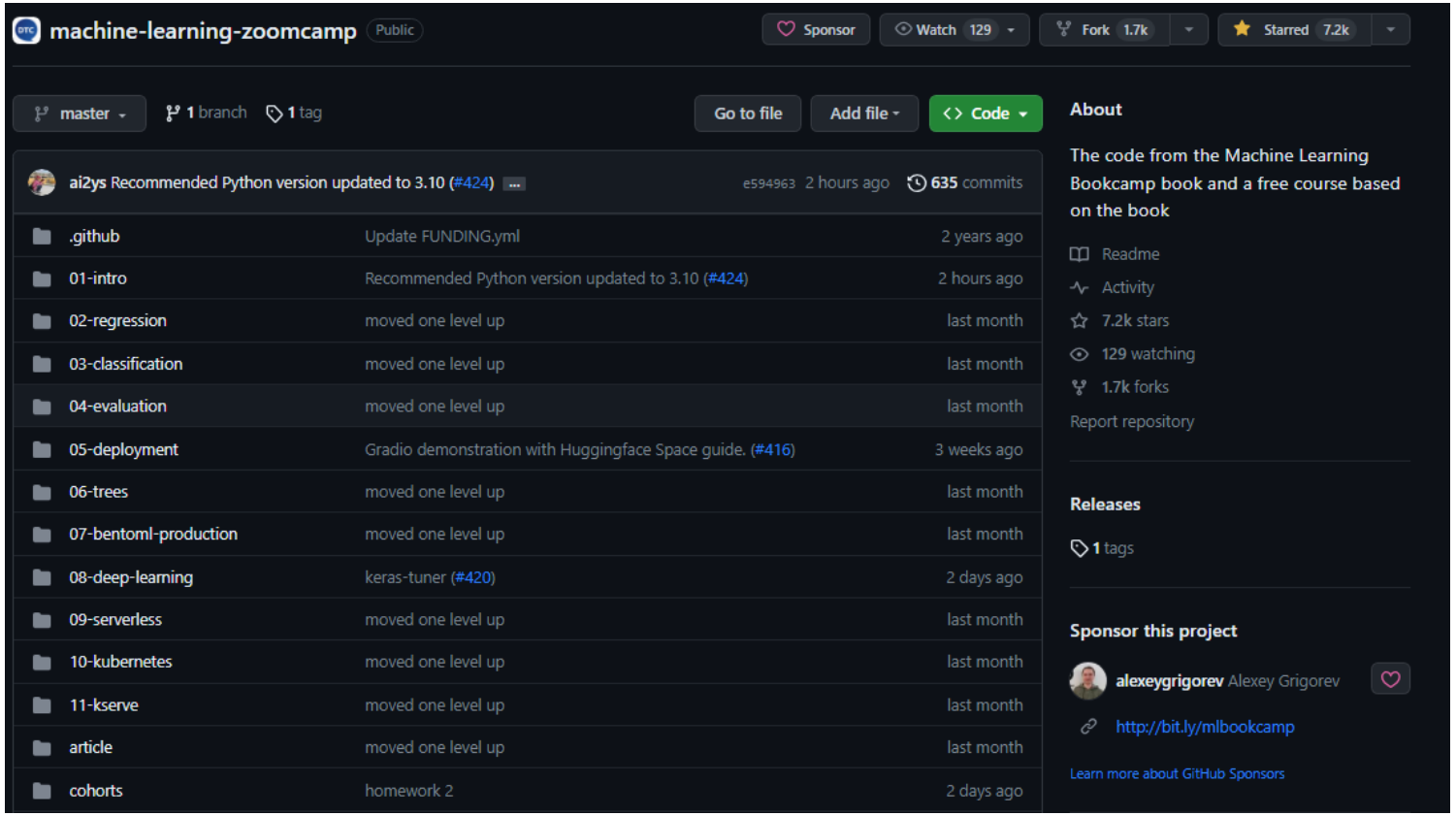
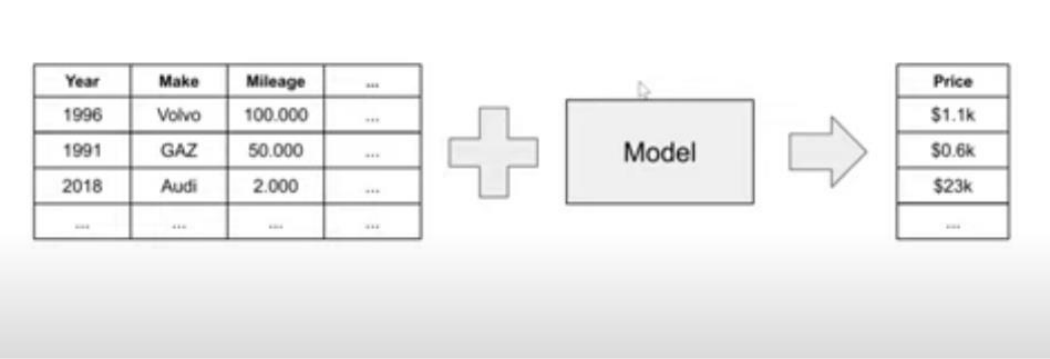
# Machine Learning Zoomcamp

Data Talks Club: [Link]



- ML is the process of extracting patterns from data
- Data is of 2 types Features and Target(Prediction from Model)





# Supervised learning

- The goal of SL is to come up with a model g which takes a feature matrix ‘x’ as input and learn to predict the output as close to the target example ‘y’

## 1. Predict a numerical value: Regression

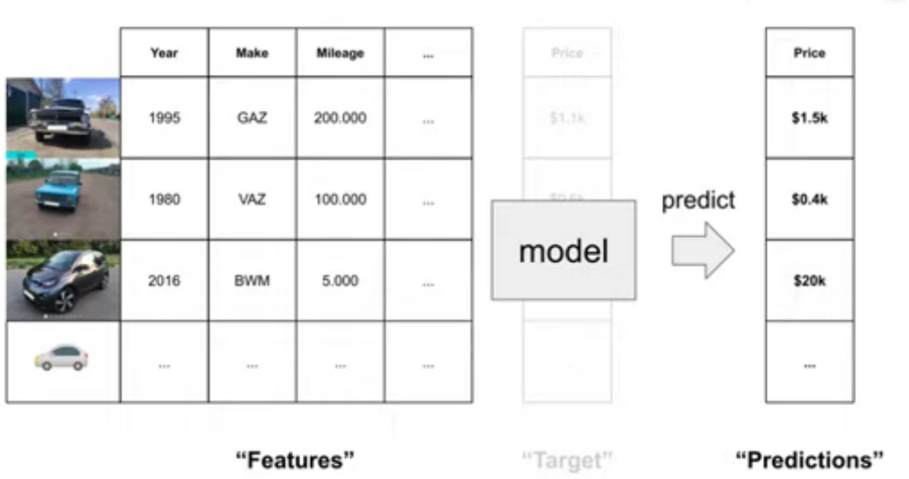
- Predicts a continuous number
- Eg: Price , Area, Height etc.

## 2. Predict a category: Classification

- Predicts a category
- Binary classification only gives one of 2 output categories
- Multicategory Classification has more than 2 options as classification output
- Eg: Spam or not, True false,

## 3. Ranking: Recommender Systems

- Rank the result based on the score assigned
- Eg: E commerce, Google search




$$g(X) \approx y$$

model FEATURES TARGET

# CRISP-DM

Cross Industry Standard Process for Data Mining by IBM

## 1. Business Understanding

- Understand the extent of problem and decide suitability of ML based system
- Define Goal
- Goal has to be measurable (Eg: reduce the spam detection by 80%)

## 2. Data Understanding

- Identify data collection and replication method
- Need to add more data or not by going back to the previous step

## 3. Data Preparation

- Clean data to be ready to apply ML models
- Tabular formatting of data with features and target variable labelling

## 4. Modelling

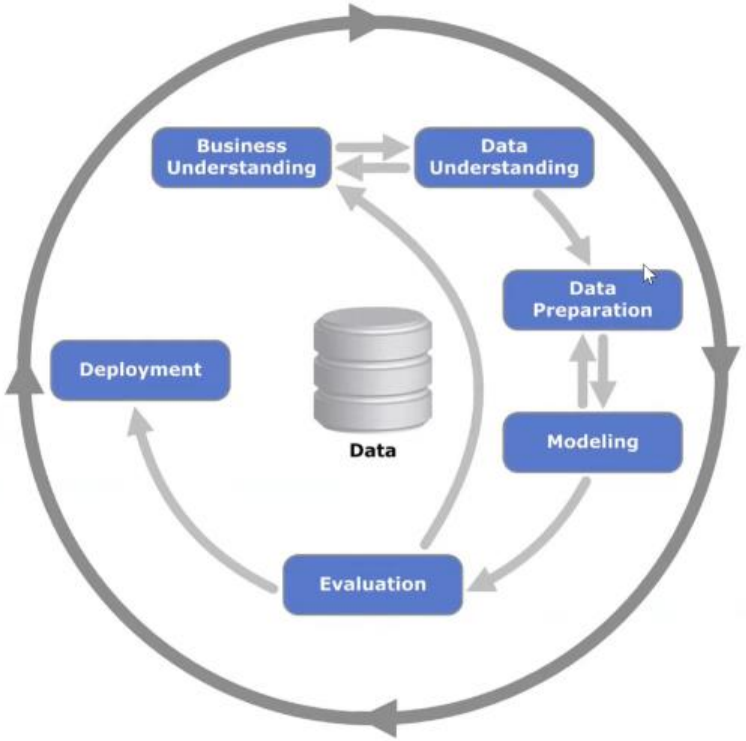
- Machine Learning Step
- If new features or problems with data detected, go back to the previous step and reformat data.

## 5. Evaluation

- Compare result with goals
- Decide whether it is acceptable

## 5. Deployment

- Deploy to users
- Test maintainability through monitoring
- Go back to business goals for further improvement



# Model Selection

## 1. Model Evaluation process:

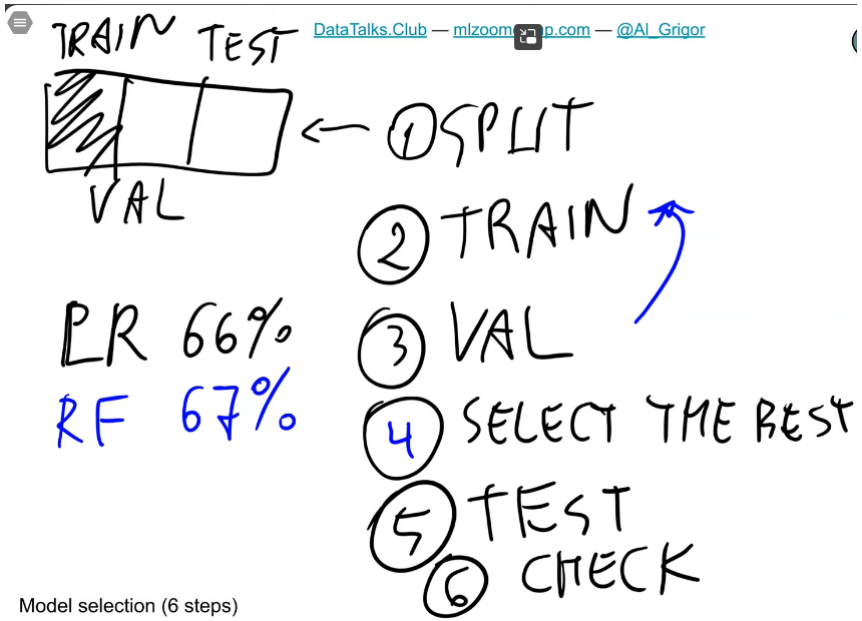
- Divide dataset into 2 parts. Training data to train the models and test data to validate the results of the models.
- In practice it might not work
- A model may luckily give good result on the validation dataset.

## 2. Multiple Comparison Problem

- When testing many models again a single validation data, one of the models may turn out lucky out of no reason.
- We divide our dataset into 3 parts instead of 2. Train data, Validate data, and finally a test data which will be used after validation to avoid the model from being lucky twice.

## 3. Solution

- After selecting a good model, retrain with 80 % of the data by combining training data with validation data to get a good final model.
- Finally use the test data on improved model to validate the results



Model selection (6 steps)

| Dataset(100%) | | | |
|---|---|---|---|
| Train 80% | Test 20% | | X |
| Train 60% | Validate 20% | Test 20% | ✓ |

# Homework

**Set up the environment**
You need to install Python, NumPy, Pandas, Matplotlib and Seaborn. For that, you can the instructions from 06-environment.md.
**Question 1**
What's the version of Pandas that you installed?
You can get the version information using the __version__ field:
pd.__version__
A 1.3.4
**Getting the data**
For this homework, we'll use the California Housing Prices dataset. Download it from here.
You can do it with wget:
wget https://raw.githubusercontent.com/alexeygrigorev/datasets/master/housing.csv
Or just open it with your browser and click "Save as...".
Now read it with Pandas.
**Question 2**
How many columns are in the dataset?
•10
•6560
•10989
•20640
**Question 3**
Which columns in the dataset have missing values?
•total_rooms
•total_bedrooms
•both of the above
•no empty columns in the dataset
**Question 4**
How many unique values does the ocean_proximity column have?
•3
•5
•7
•9

**Question 5**
What's the average value of the median_house_value for the houses located near the bay?
•49433
•124805
•259212
•380440
**Question 6**
1.Calculate the average of total_bedrooms column in the dataset.
2.Use the fillna method to fill the missing values in total_bedrooms with the mean value from the previous step.
3.Now, calculate the average of total_bedrooms again.
4.Has it changed?
Has it changed?
Hint: take into account only 3 digits after the decimal point.
•Yes
•No
**Question 7**
1.Select all the options located on islands.
2.Select only columns housing_median_age, total_rooms, total_bedrooms.
3.Get the underlying NumPy array. Let's call it X.
4.Compute matrix-matrix multiplication between the transpose of X and X. To get the transpose, use X.T. Let's call the result XTX.
5.Compute the inverse of XTX.
6.Create an array y with values [950, 1300, 800, 1000, 1300].
7.Multiply the inverse of XTX with the transpose of X, and then multiply the result by y. Call the result w.
8.What's the value of the last element of w?
**Note**: You just implemented linear regression. We'll talk about it in the next lesson.
•-1.4812
•0.001
•5.6992
•23.1233