Learning Objectives

The ABC of Machine Learning

Imbalanced Dataset

One-hot encoding

Session Details

DPhi

The ABC of Machine Learning

DPhi

What is Machine Learning?

- Machine learning provides systems the ability to **automatically learn and improve from experience without being explicitly programmed**.

- It allow computers to discover hidden and useful insights

- **In nutshell, Machine Learning is a new way of communicating your wishes to a computer.**

Machine Learning is used in..

- **Fraud detection** - **Eg:** Credit card fraud detection. It will help us to detect whether a transaction is fraud or not.

- **Email spam filtering** - **Eg:** Helps in categorising whether a particular email should go in inbox or spam box.

- **Recommendation engines** - **Eg:** E-commerce platforms like Amazon can recommend you a similar product based on your previously browsed list of products

- and many more!!!

What is a Machine Learning (ML) model?

- For now, let's consider it is a Magical box that help us to predict what we want. In the below case we want to predict whether an incoming email should land in our inbox or spam box. We will discuss more about ML models soon.

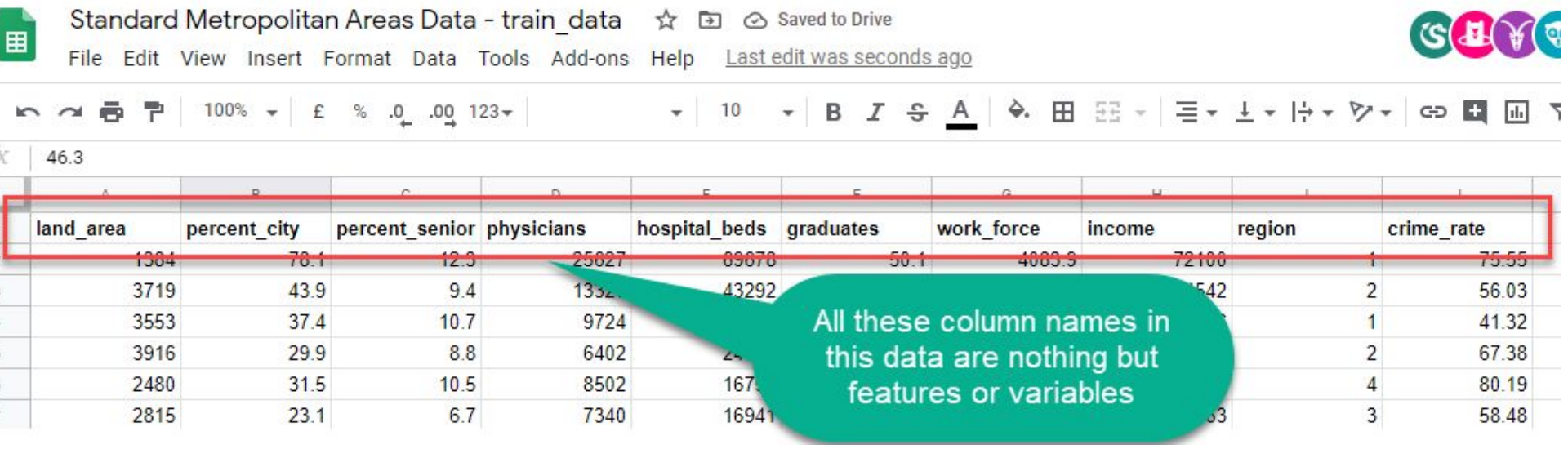
In other terms this is nothing but **data**. This data will have variables such as: sender email id, subject of email, email body etc

Once the incoming emails go through the Machine Learning Model it categorizes and predicts whether a mail should go in your inbox or spam box

Variables/features

- **Features or Variables:** These are the the most common terms that we would come across from now on.

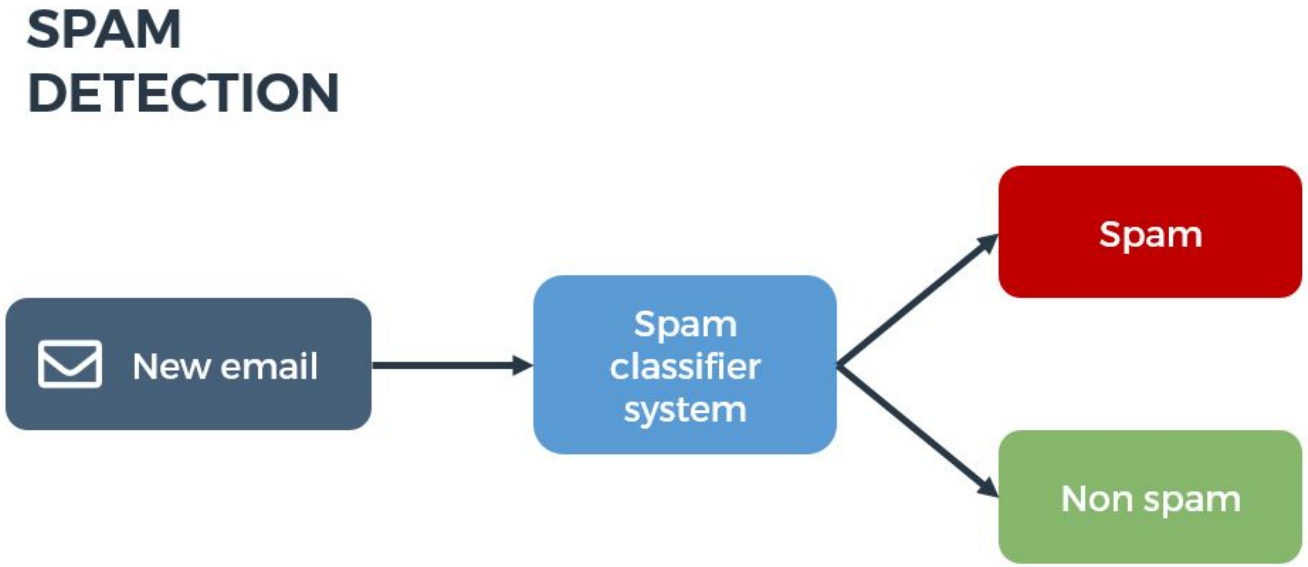
- **Features and Variables both are the same in a dataset**, they are often interchangeably used. So there is no need to worry about it!



DPhi

Target/Label Variable

- The target variable or label of a dataset is the feature of a dataset about which you want to gain a deeper understanding.
- It is the variable that is, or should be the output.
- In the example of detecting spam emails, the label will be the category the email belongs to, i.e it will be either ‘spam’ or ‘not spam’.



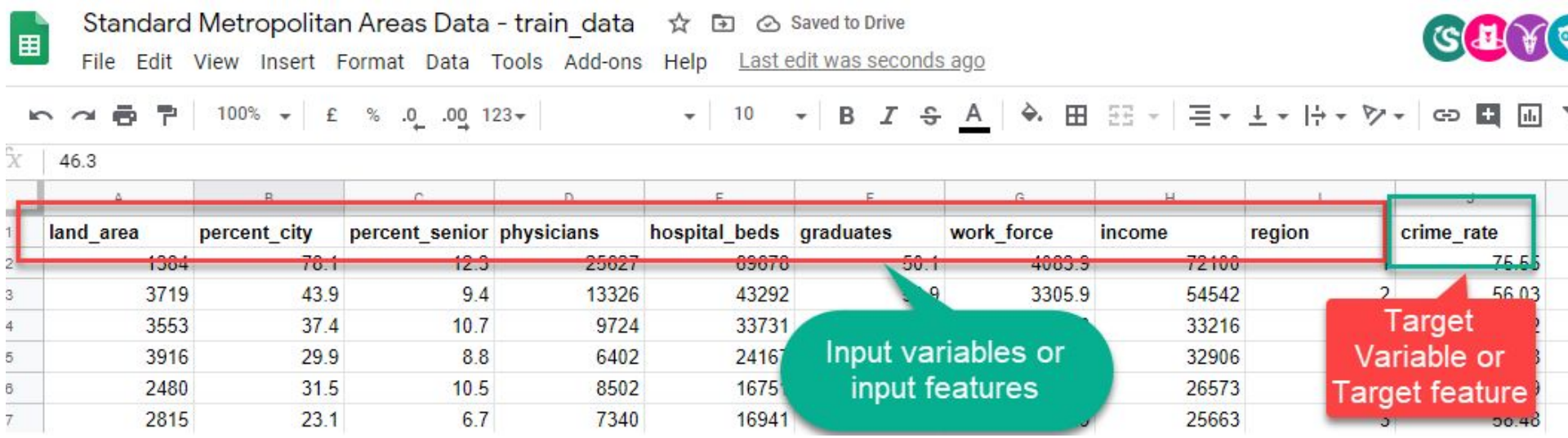
Predictor/Input Variables

- One or more variables that are used to determine (or predict) the 'Target Variable' are known as Input Variables. They are sometimes called Predictor Variable as well.

- In the spam detector example, the features could include the following:
  - words in the email text
  - sender's address
  - time of day the email was sent
  - email contains the phrase "congrats you won $1 billion - share your bank details."

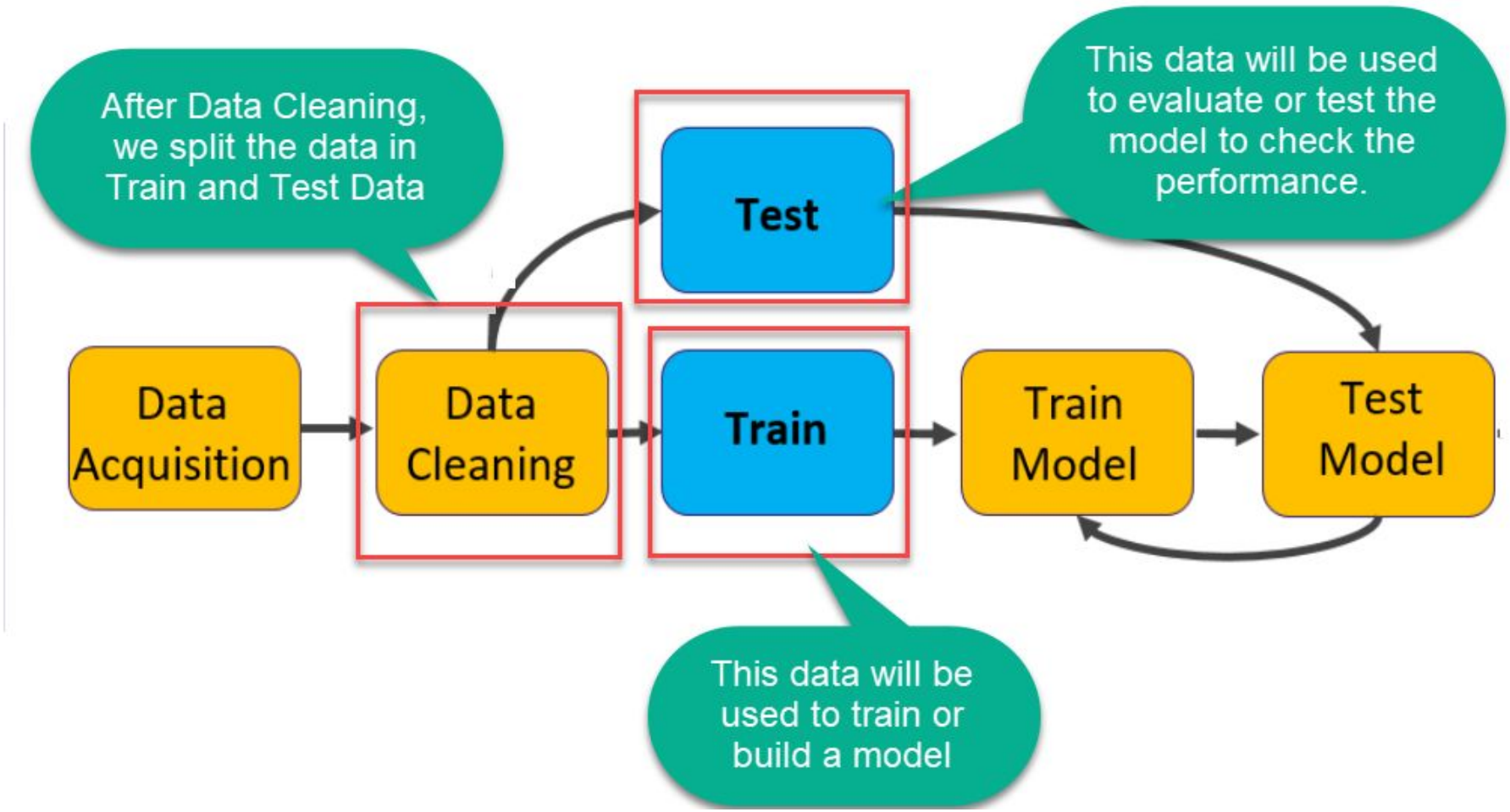
Target and Input variables

- Remember the Standard Metropolitan Areas Data used in previous slides? In that dataset **we might be curious to predict "crime_rate" in future**, so that becomes our target variable and rest of the variables become input variables for building a machine learning model.



DPhi

Train and Test Set



Train and Test Set

- To use an analogy, let's say you teach a child to multiply by letting the kid train on the small multiplication table, i.e. everything from 1*1 to 9*9.

- Next, you test whether the kid is able to perform the same multiplications. The result is a success. The kid gets it right almost every time.

- What's the problem here?

- You don't know if the kid understands multiplication at all, or has simply memorized the table!

DPhi

Train and Test Dataset

- So what you would do instead is test the kid on multiplications like 11*12, that are outside of the table.

- This is exactly why we need to test machine learning models on **unseen data or test data.** Otherwise, we have no way of knowing whether the algorithm has learned a generalizable pattern or has simply memorized the training data.

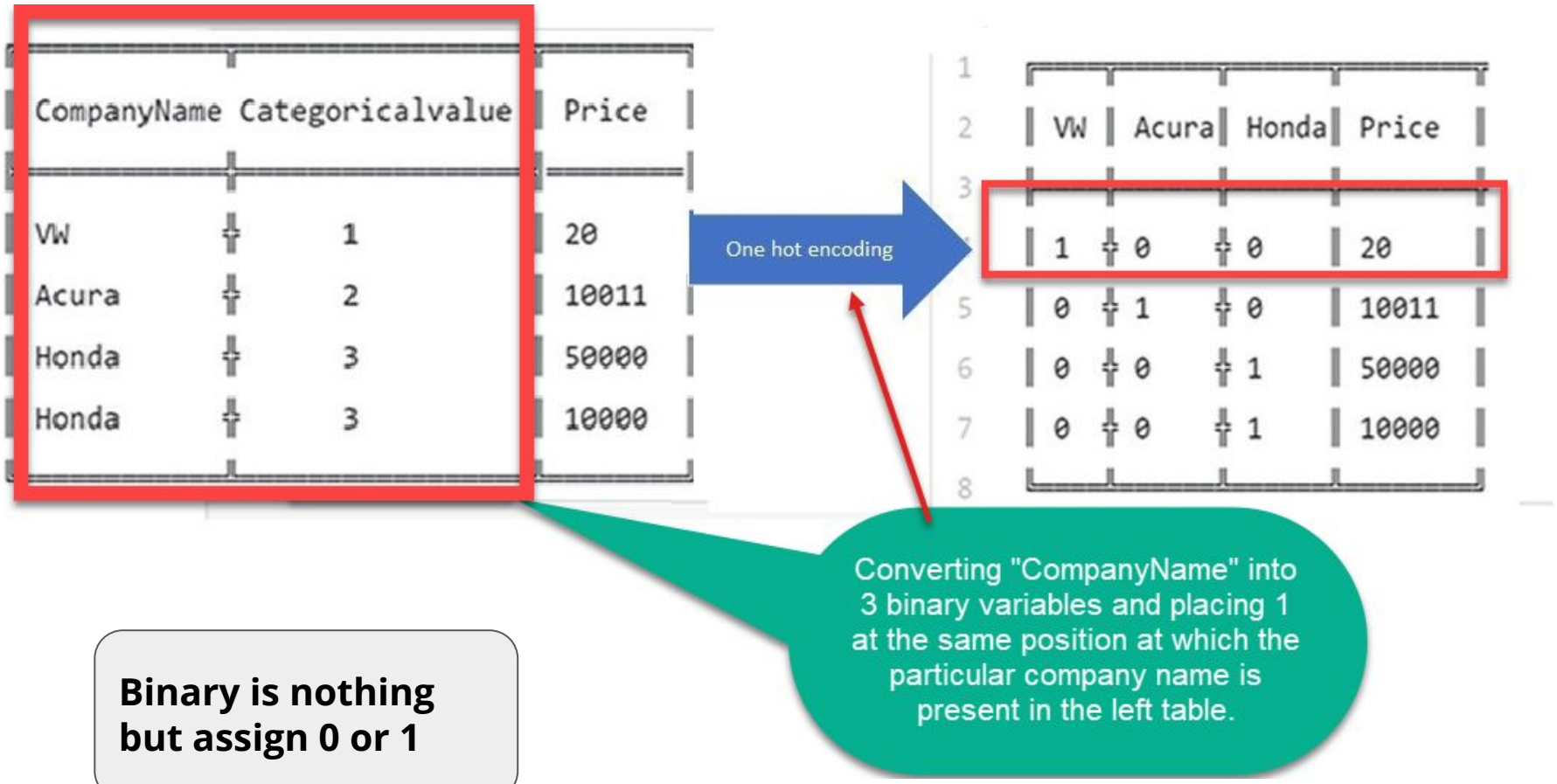
- **TRAINING DATA:** The observations in the training set form the experience that the algorithm uses to learn.

- **TEST DATA:** The test set is a set of observations used to evaluate the performance of the model using some performance metric. It is important that no observations from the training set are included in the test set. If the test set does contain examples from the training set, it will be difficult to assess whether the algorithm has learned to generalize from the training set or has simply memorized it.

Train Test Split

- Consider an Example where our Original Dataset has 1000 rows.
- When we start building our ML model we will split our datatset into two parts (70% train data and 30% test data).
- We will train our model on 70% of data i.e 700 rows and then test our model performance on 30% of data i.e 300 rows. As discussed above while testing our model we will not provide the outcome to our model for the test data although we know the outcome and instead let our model give us the outcome for those 300 rows.
- Later we will compare the outcome of our model to the original outcome of our test data to get the accuracy of our model predictions.
- For splitting our data to training and testing set we use **train_test_split method of scikit-learn library.**

One Hot Encoding



DPhi

One Hot Encoding

- One hot encoding is a process by which categorical variables (do you remember what categorical variables mean from the stats module?) are converted into a form that could be provided to ML algorithms to do a better job in prediction.
- Let's consider the dataset in the previous slide
- The Categoricalvalue is just a number assigned to each car company name.
- The problem with this is that it assumes higher the categorical value, better the category. But that's definitely not the case, right?
- This is why we use one hot encoder to perform "binarization"(representation in 0 and 1) of the category and include it as a feature to train the model.

One Hot Encoding

- Now what do the 0s and 1s actually represent?

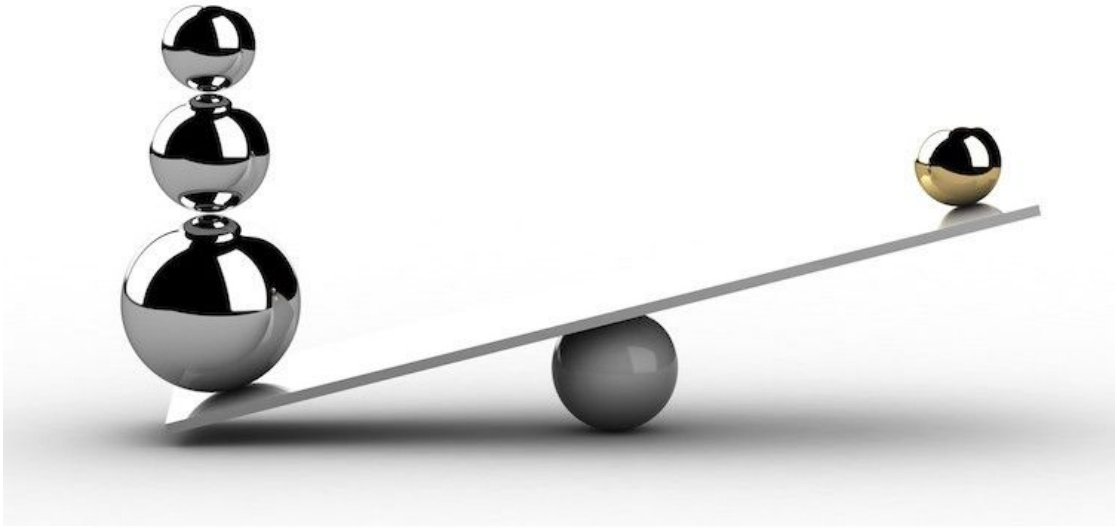
- Look closely at the table on the right in previous slide. In the company name variable example, there are 3 categories (3 car companies) and therefore 3 binary variables are needed (VW, Acura, Honda).

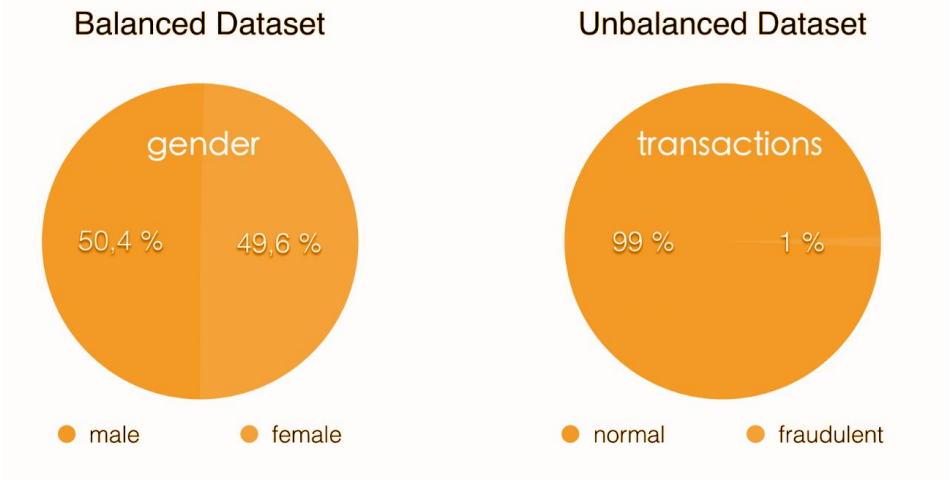
- A “1” value is placed in the binary variable at the same position at which the particular company name is present in the left table. The rest are kept “0” in that variable’s column. The remaining 2 variable columns are filled in similar manner.

- In short, all the elements of the vector are 0 except one, which has 1 as its value.

- MUST READ: The following resource justifies the necessity of one hot encoding our data: https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/

Imbalanced Dataset

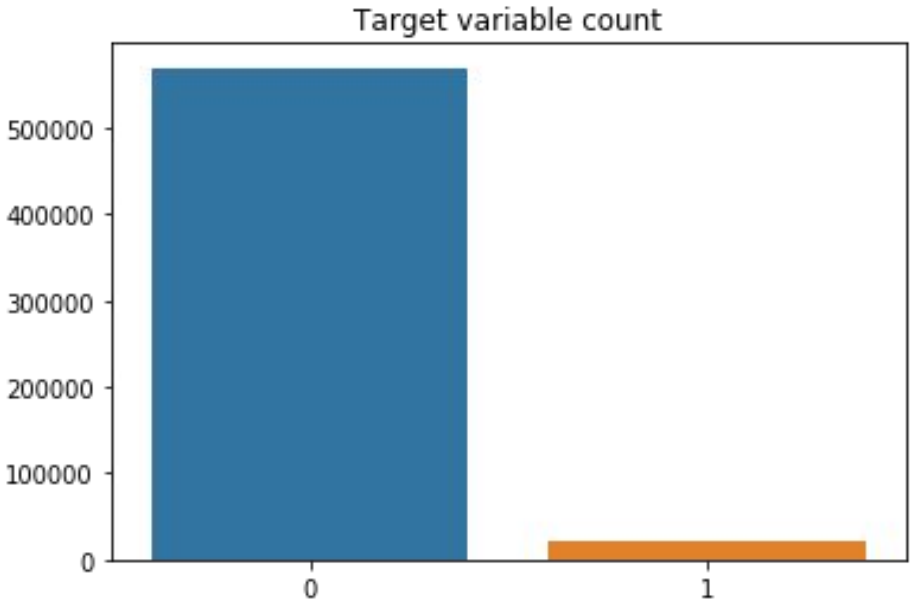




DPhi

Imbalanced Dataset

- For example, suppose you have a credit card transaction data and you are supposed to predict fraudulent transactions. You'll likely have 10,000 authentic transactions for every 1 fraudulent transaction, that's quite an imbalance!
- **In machine learning terms:** Often you'll have a large amount of data/observations for one class (referred to as the majority class), and much fewer observations for one or more other classes (referred to as the minority classes).



DPhi

Imbalanced Dataset

- The problem is that machine learning models trained on unbalanced datasets often have poor results when they have to generalize (predict a class or classify unseen observations). Despite the algorithm you choose, some models will be more susceptible to unbalanced data than others. Ultimately, this means you will not end up with a good model, and the reasons include:

    - The algorithm receives significantly more examples from one class, prompting it to be biased towards that particular class.
    - It does not learn what makes the other class “different” and fails to understand the underlying patterns that allow us to distinguish classes.

Imbalanced Dataset

Can you think of any other example where there is class imbalance?

Data Pre-Processing

- Instructors' recommended article: https://towardsdatascience.com/data-preprocessing-concepts-fa946d11c825

P.S: Try opening in incognito if it is asking for a premium subscription upgrade

Download Dataset

- **Download IEEE Fraud Dataset:**
  https://bit.ly/Dataset_FraudDetection

!!! - it is a huge dataset so opening it in excel might be difficult, so please open it via python environment - (Colab or Jupyter notebook)

- Read about this dataset here:
  https://www.kaggle.com/c/ieee-fraud-detection/data

DPhi

About the Dataset

IEEE Fraud Dataset

- The data is broken into two files identity and transaction, which are joined by TransactionID

Transaction Table *

- **TransactionDT:** timedelta from a given reference datetime (not an actual timestamp)
- **TransactionAMT:** transaction payment amount in USD
- **ProductCD:** product code, the product for each transaction
- **card1 - card6:** payment card information, such as card type, card category, issue bank, country, etc.
- **addr:** address of the customer
- **dist:** distance
- **P_ and (R__) emaildomain:** purchaser and recipient email domain

About the Dataset

- **M1-M9:** match, such as names on card and address, etc.

- Categorical Features:
  - ProductCD
  - card1 - card6
  - addr1, addr2
  - Pemaildomain Remaildomain
  - M1 - M9 (bank sensitive data)

Note: Some of the feature/variable description is not given as

About the Dataset

Identity Table *

- Variables in this table are identity information – network connection information (IP, ISP, Proxy, etc) and digital signature (UA/browser/os/version, etc) associated with transactions.
- They're collected by Vesta’s fraud protection system and digital security partners.
- (The field names are masked and pairwise dictionary will not be provided for privacy protection and contract agreement)

- Categorical Features:
  - DeviceType
  - DeviceInfo
  - id12 - id38

Notebook Link

https://github.com/dphi-official/Data_Science_Bootcamp/blob/master/Week2/Introduction_to_Imbalanced_class.ipynb

!! This dataset takes a lot of time to upload on colab. You may want to consider using Jupyter Notebook

DPhi

That’s it for this unit. Thank you!

Feel free to post any queries on Discuss.

DPhi