DATA PRE-PROCESSING

By Bharat Ram Ammu
Data Scientist,
Complidata,
Belgium
ammubharatram@gmail.com
Linkedin Profile:
https://www.linkedin.com/in/bharatramammu

Topics today – Data Pre-processing

- Handling Missing Values
- Handling Imbalanced datasets, Oversampling - SMOTE
- Standardization/Normalization and transformation for data
- Explained with an example on Detecting Fraudulent Transactions in a given dataset!

Pre-processing – what, why and how?

What?

Pre-processing is the process of preparing the data for training.

Why?

- Data is not ready-made for us:

✔ Missing values

✔ Wrongful Data entries

✔ Class Imbalance

✔ Different scales of data..

How?

- Handling Missing Values
- Handling Imbalanced datasets, Oversampling - SMOTE
- Standardization/Normalization and transformation for data

Missing values in data

Most datasets are not perfect, they have missing values.

Missing values can be due to:

- Missed entries by participants (in case of surveys)
- Missed information by database managers
- That variable is not relevant for that data point

(e.g. in a customer dataset for banks, ‘No of children’ not relevant for unmarried customers. Hence, results in an NA value.)

Dataset	Observations	True positives	No of variables	% of fraud cases
IEEE	57049	2005	180	3.5

Example dataset: IEEE Fraud Dataset- a 10% sample used for simplicity

IEEE Fraud Dataset was provided at Kaggle a year ago:

- **Categorical Features – Transaction**
  - ProductCD – Product code
  - card1 - card6 : payment card information, such as card type, card category, issue bank, country, etc.
  - addr1, addr2
  - P_emaildomain – Purchaser
  - R_emaildomain- Recipient
  - M1 - M9 – Match between names on card and address etc.
- **Categorical Features - Identity**
  - DeviceType
  - DeviceInfo
  - id_12 - id_38
  - The TransactionDT feature is a timedelta from a given reference datetime (not an actual timestamp).
- **Outcome/Target Variable – isFraud** –
  - whether transaction is fraud or not

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

Ideas for Missingness Mechanisms

Missingness in dataset can affect a machine learning problem in different ways :

1. Missing completely at random

(eg: forgot to fill in survey, forgot to enter by data entry etc)

2. Missingness is related to other variables(predictors) used for prediction.

(E.g: Fraud dataset, ‘VAT number’ doesn’t exist because customer is of type ‘individual and not business. Here ‘customer type’ is another predictor variable for fraud.

Ideas for Missingness Mechanisms

3. Missingness is related to outcome variable predicted itself and is hence not random.

(E.g: Fraud dataset, ‘Transaction message’ doesn’t exist because customer is doing perhaps something suspicious. Here, imputation might need advanced techniques. Beyond the scope of this course..

Source: Tim Bock on DisplayR

Why fill missing values?

- How should a model read a missing value?
- Doesn’t it change what machine learns if it reads it as 0?
- Mathematical models cannot understand what a missing value means.

Techniques for filling missing values

- Single Imputation
- Regression Imputation
- Multiple Imputation

Which variables to impute-
filling missing values vs using unreal information

- Imputation or filling in unreal values doesn't come without consequences.
- Is it worth using variables which have 80% missing values which have been imputed?
- Typical recommendation for variables to be used are variables with < 20% missingness
- Heuristic for % of missingness ranges from 10% to 20% and hasn't reached consensus in academic research. (Schlomer, Bauman, and Card 2010)
- Conclusion: In our example, we will just impute variables with less than 20% missingness and leave out variables with missingness higher than that

Simple Imputation and how it helps

Simple Imputation: replace the missing value in a variable by the mean/mode of the variable:

If variable is numeric: replace with mean of the variable

If variable is categorical: replace with mode of the variable

How it helps?

Imputing by mean or median balances the data distribution

Code for imputation: Example

Other Imputation Techniques

- Multiple Imputation: Multiple Imputation fills in estimates for the missing data. But to capture the uncertainty in those estimates, MI estimates the values multiple times.

- Example In Python: Iterative Imputer

A strategy for imputing missing values by modeling each feature with missing values as a function of other features in a round-robin fashion.

Example of Multivariate feature imputation:

Iterative Imputer

Recommended Imputations for various Missingness Mechanisms

1. Missing fully at random - reasonably good to use single imputation techniques (by mean for numeric variables and mode for categorical variables).
2. Missingness is related to other variables(predictors) used for prediction- Maybe use multiple imputation here

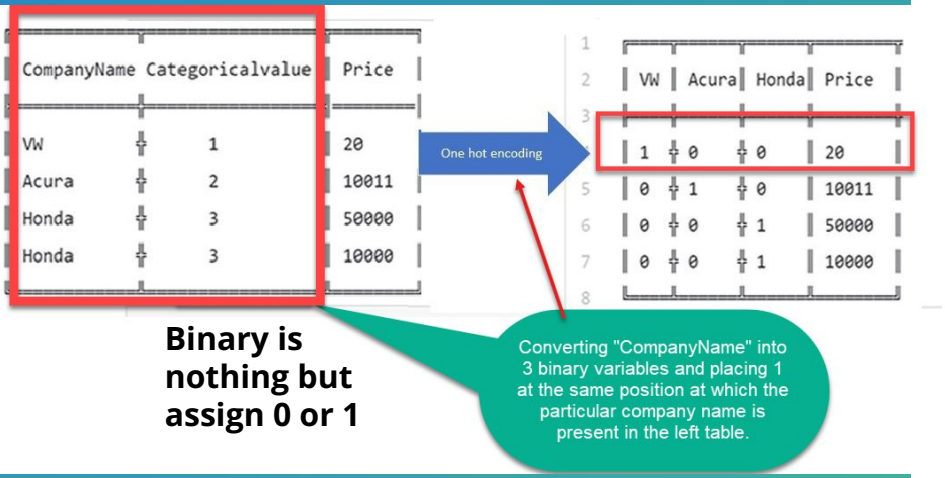
3. Missingness is related to outcome variable predicted itself - Here, imputation might need advanced techniques. Beyond the scope of this course.

References

- Common Methods to Imputing Missing Data

https://www.theanalysisfactor.com/seven-ways-to-make-up-data-common-methods-to-imputing-missing-data/

- Multiple Imputation:
- https://www.theanalysisfactor.com/missing-data-two-recommended-solutions/
- Imputation in Python using scikit-learn:
- https://scikit-learn.org/stable/modules/impute.html
- Schlomer, Gabriel L., Sheri Bauman, and Noel A. Card. 2010. “Best Practices for Missing Data Management in Counseling Psychology.” *Journal of Counseling Psychology*.

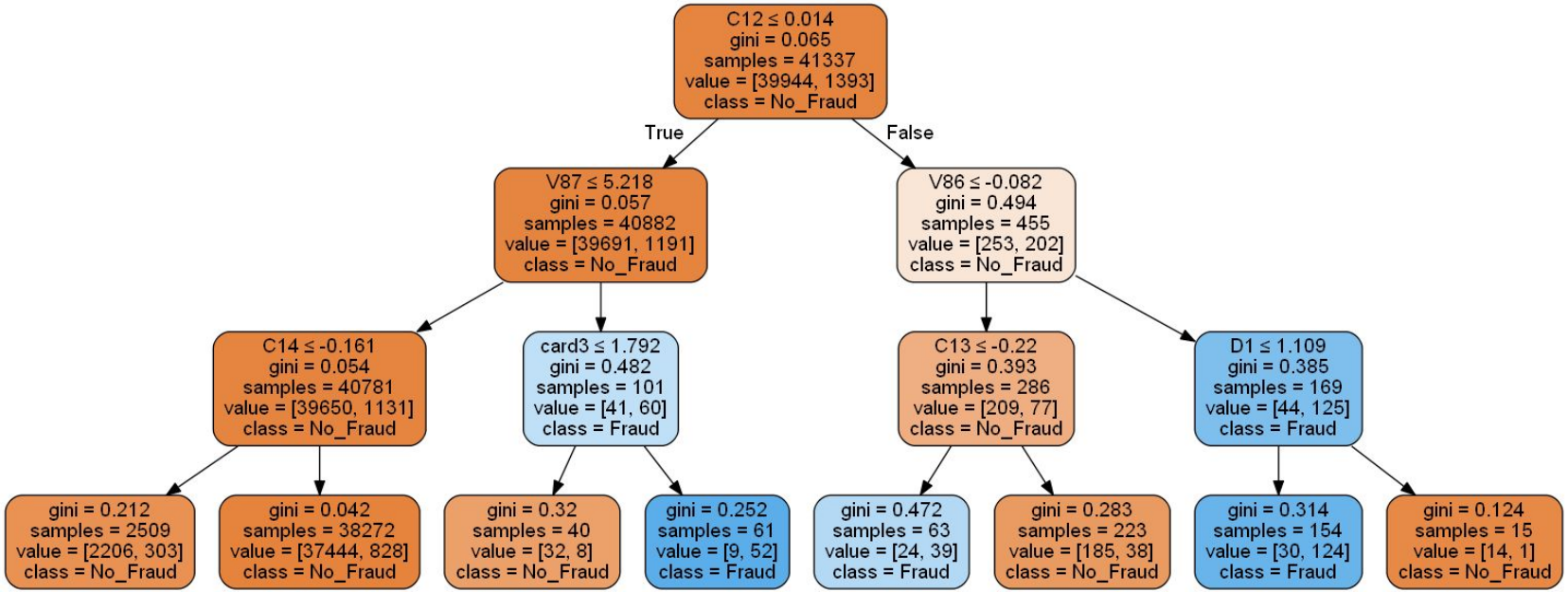


One Hot encoding in Python

- In Python, typically machine learning models expect inputs to be numbers.
- Hence, it is easier to convert all categorical variables to numeric.
- One such techniques popularly used is one-hot encoding.
- One-hot encoding is checking
- Example :

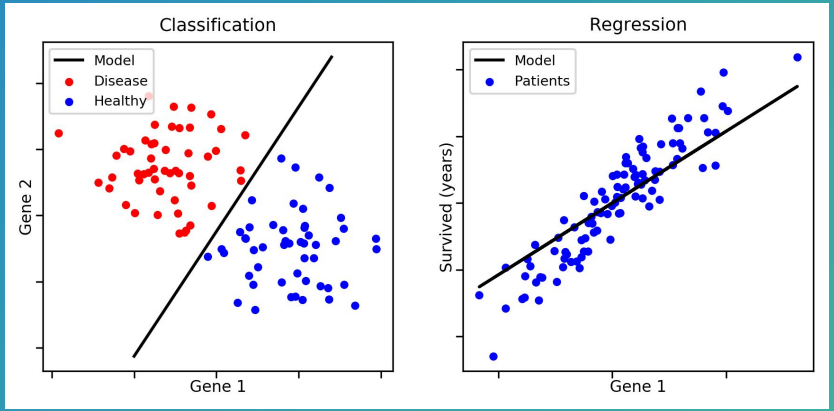
DEMONSTRATE THE FIRST DECISION TREE

A simple decision tree model



CLASS IMBALANCE PROBLEM

Machine Learning : Classification Vs Regression



Classification vs Regression:

- **Classification**

✔ Classification is the problem of classifying the outcome into two or more outcomes

- Classification examples:

✔ Predict colour: Red/Blue/Green etc

✔ Predict if name matches/not

- **Regression**

✔ Regression is the problem of predicting a continuous outcome (a numeric outcome)

- Regression examples:

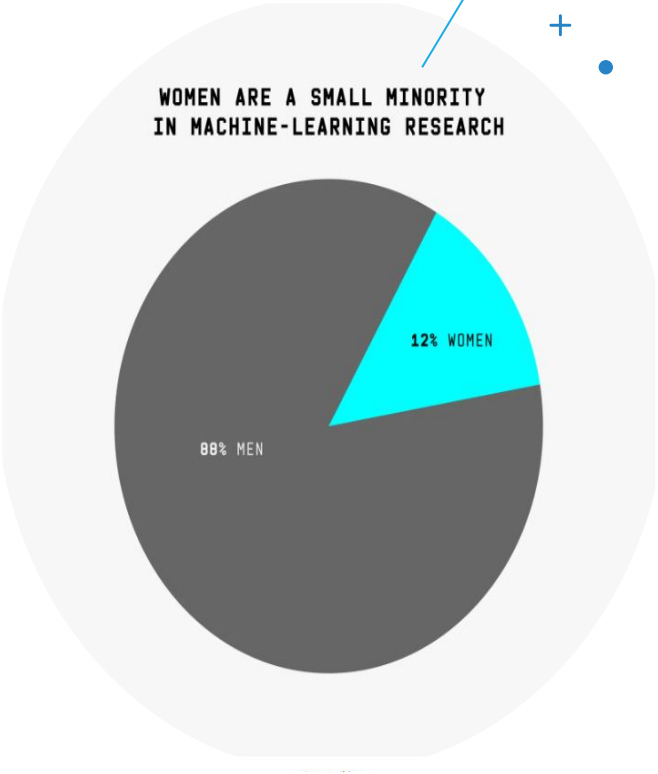
✔ Predict price of house

✔ Predict age of a dog based on data of dog images!

The problem of Class Imbalance

- Classification Problem:
- Class Imbalance: Minority class constitutes a very minute fraction compared to majority class.
- Why?
- Class imbalance in classification problem is too less data points of one class compared to another class we are trying to predict.
- How it affects?
- It leads to machine to learn too much of the dominant class and too less about the minority class!
- Examples:
- Millions of black people affected by racial bias in health-care algorithms (Source: Nature)
- Credit card approval algorithms may be biased against women. (Source: Globalnews.ca)

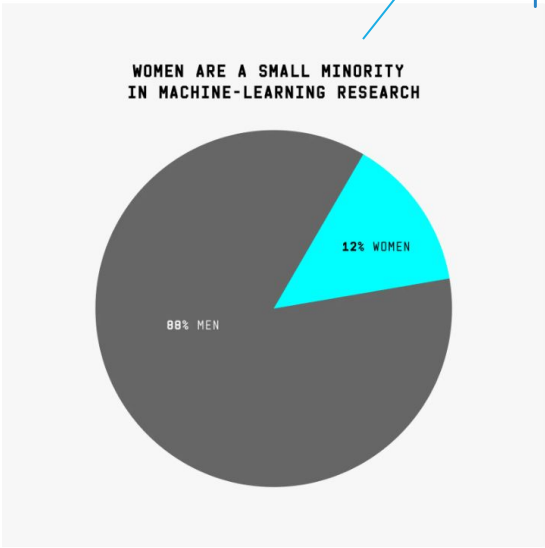
Will this lead to bias against women?



Example of Class Imbalance

- Given a database of machine learning publications, if the problem is to predict whether a researcher is male or female, will the default prediction be biased by machine learning?

Will this lead to bias against women?

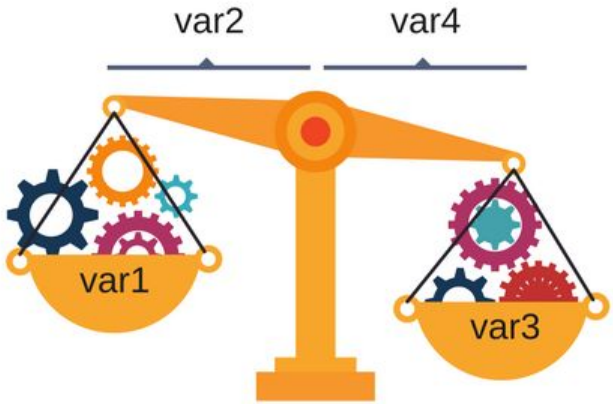


When is class imbalance a problem?

- Class imbalance is a problem when there are too less minority class (fraud) observations for model to learn from.
- One needs to decide when to create new minority class (Fraud) observations or remove existing majority (normal transactions) class observations.

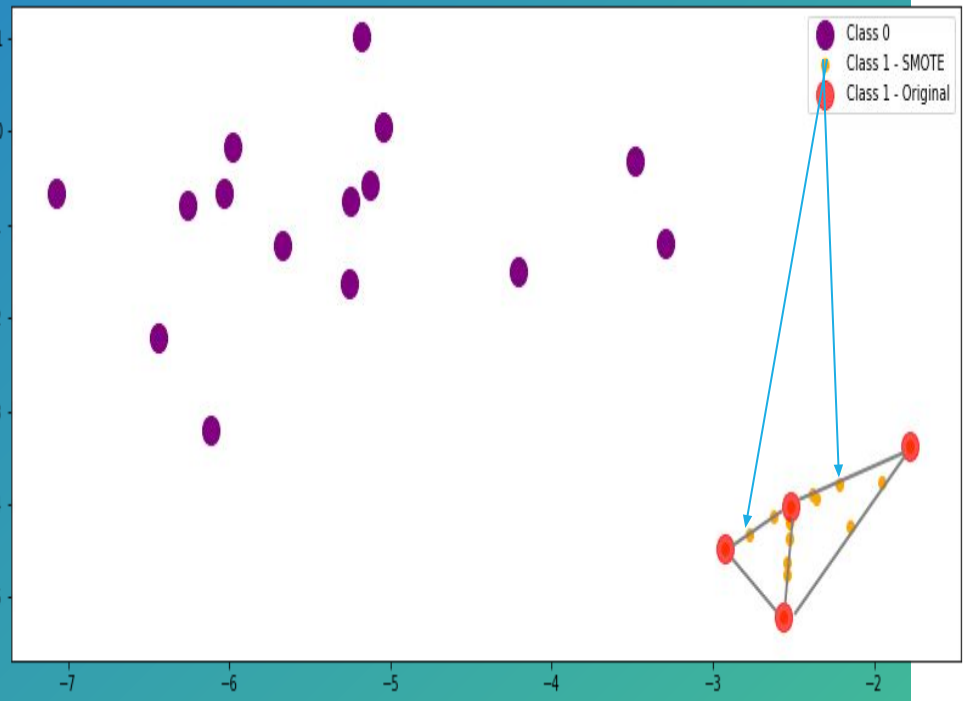
Class Imbalance in Machine learning : in our example: Balanced Scale Data

- Majority class:
- Minority class:



How to Handle Imbalanced datasets

- Oversampling: Artificially increase the minority class (eg: by duplicating minority class observations, artificially creating them...)
- Under sampling: Sample out observations from majority class to have a certain ratio between majority and minority class
- SMOTE: Combination of both.



Delving deeper into one technique- SMOTE

- SMOTE:

Synthetically (S) creating minority (M) class observations leading to oversampling (O) using this technique (TE) and under sampling majority to get a certain ratio between the classes.

- Proposed by Chawla et al 2002. (ref)

PERFORMANCE WITH AND WITHOUT HANDLING CLASS IMBALANCE

Performance with and without handling class imbalance

- AUC (a performance score for decision tree classifier) is slightly better using the “SMOTE’d” data based model.
- We can play around with parameters in SMOTE and further improve the model.
- We can also use advanced machine learning models to improve further!

References

- SMOTE example in Python: https://imbalanced-learn.readthedocs.io/en/stable/generated/imblearn.over_sampling.SMOTE.html
- SMOTE original paper:

https://arxiv.org/pdf/1106.1813.pdf

- Balanced Scale Data:

http://archive.ics.uci.edu/ml/machine-learning-databases/balance-scale/

- Example used to demonstrate data:

https://elitedatascience.com/imbalanced-classes

Standardization/scaling/Normalization of data - what, why and how?

What?

- Standardization/Scaling is bringing all variables used for building model to the same scale

Why?

- It balances the overeffect of variables with higher range (let us example in next slide)
- Sometimes, it also helps in speeding up the calculations in an algorithm.
- It is important for techniques which use distance metrics.

How?

- **Scale**– It means to change the range of values but without changing the shape of distribution. Range is often set to 0 to 1.
- **Standardize** means changing values so that distribution standard deviation from mean equals to one,output will be very close to normal distribution.
- **NORMALIZE**-It can be used either of above things

Why scaling- in our example

- Let's say you have two input vectors: X1 and X2. and let's say X1 has range(0.1 to 0.8) and X2 has range(3000 to 50000). Now your SVM classifier will be a linear boundary lying in X1-X2 plane. My claim is that the slope of linear decision boundary should not depend on the range of X1 and X2, but instead upon the distribution of points.

Various scaling methods

- Min-Max Scaler
- Robust Scaler
- Standard Scaler
- Normalizer

When to scale data?

- If you build models using scaled data, it may require scaling back to original variables to interpret variables' effect on outcome predicted.

Put all the pre-processing techniques together

- Handle missing values via:

 Imputation:

1. Single Imputation
2. Multiple Imputation

- Handle class imbalance:

1. Other techniques – Oversampling, undersampling
2. SMOTE

- Additional: Standardization/scaling/Normalization of data

References

- Why scaling is important and techniques:

https://mc.ai/why-scaling-is-important-in-machine-learning/

Slide Download Link

You can download these slides from the below link:

https://docs.google.com/presentation/d/10hLbtd-xIaUx0RPulzf3139z44Wb1jEpIpEw0nBpDEg/edit?usp=sharing

THANK YOU FOR YOUR ATTENTION! WISH YOU HAPPY LEARNING!