Anomaly Detection with Bayesian Neural Networks

Theodore Ladas

T.Ladas@sms.ed.ac.uk

THE UNIVERSITY *of* EDINBURGH
School of Mathematics

Motivation

In General

- Anomaly detection is used in the military, for cybersecurity and in banking.
- The project aims to automate fraudulent transactions flagging.
- Using Bayesian ML methods to train a model to predict an outcome of interest
- Big emphasis is also given in the connection between EDA and Model Explanation

Data

- The dataset is very well-known real-life dataset for prototype creation.
- Matrix of 5000 (rows) $\times$ 12 (columns), where
  - rows $\rightarrow$ Wines
  - columns $\rightarrow$ Features, such as degrees of *Alchohol*, wine *Acidity* etc.

Software

- R - for Exploring the dataset (libraries: ggplot2, tidyverse).
- Python - for Building the model (libraries: keras, tensorflow, shap).

Exploration

PCA Explanation

- The most important exploration of the data is a Principal Components Analysis.
- Singular value decomposition of X produces three matrixes, U ,S and V.
- The V matrix, shows how much each feature contributes to each PC.

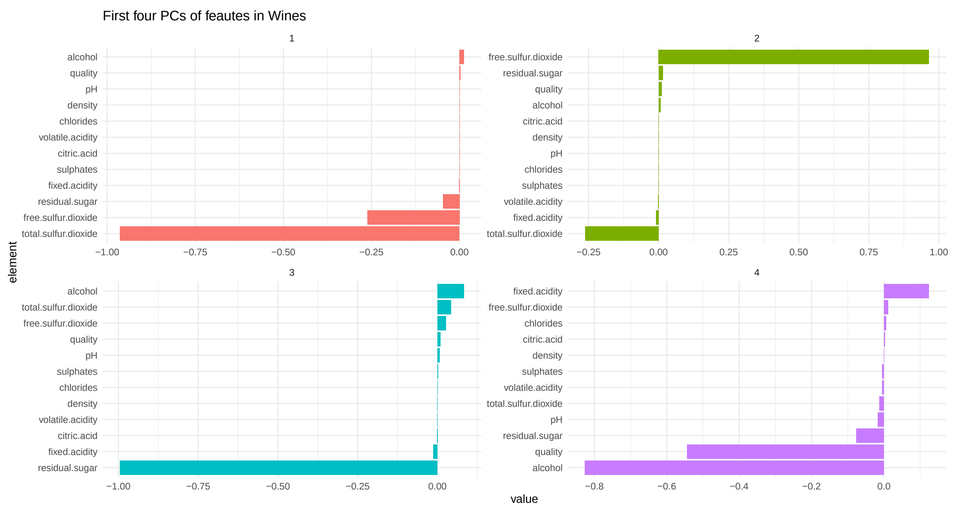
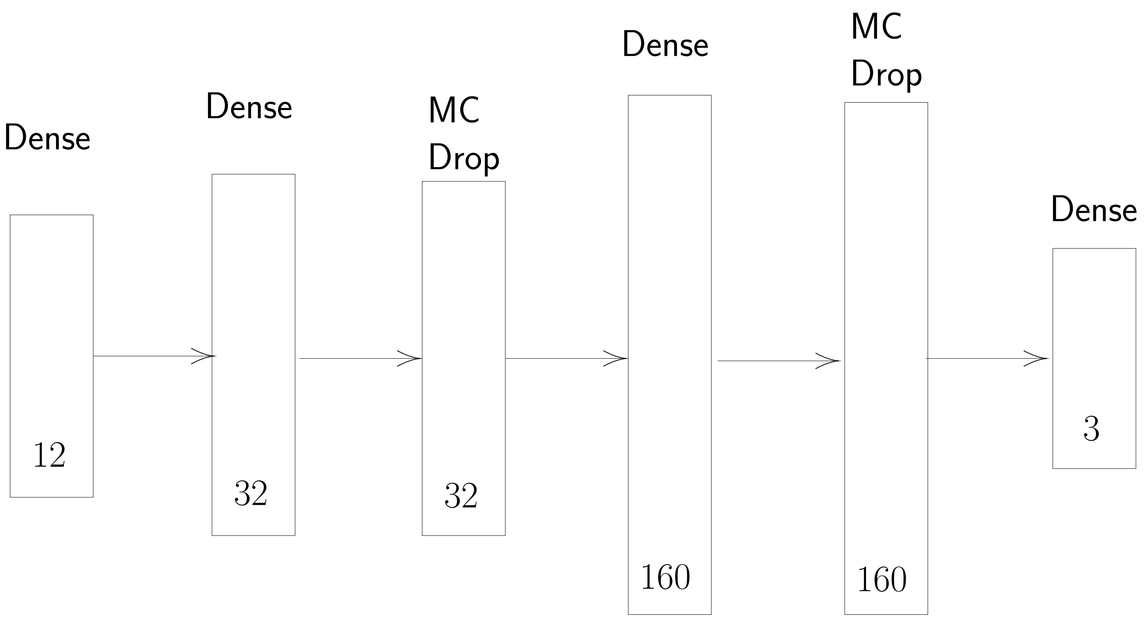


Figure 1: First four principal component loading scores per feature

Model

Monte Carlo Dropout BNN



Dropout layer

- Dropout layers work only during training.
- They randomly ignore a percentage of neurons and their connections.
- Randomly means, by random draws of a Bernoulli distribution.
- This process generates a **deterministic** prediction.

Monte Carlo Dropout layer

- MC dropout is a wrapper of the simple dropout function.
- It preserves the random dropout of neurons at testing time.
- This process generates a **stochastic** prediction.

Validation

Explanation of model fitting

- There exist a lot of variance in the prediction due to randomness.
- The validation and the training accuracy are at the same level.

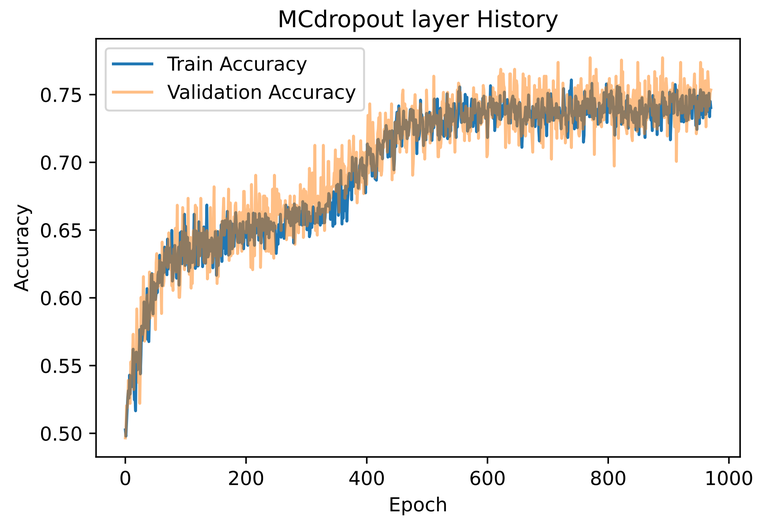


Figure 2: Epochs training

Explanation

Shap Values

- The most influential features in descending order.
- Further granularity, fixed acidity low, then low impact on prediction class 1.
- Prior knowledge gained in EDA is confirmed in model explanation by SHAP.

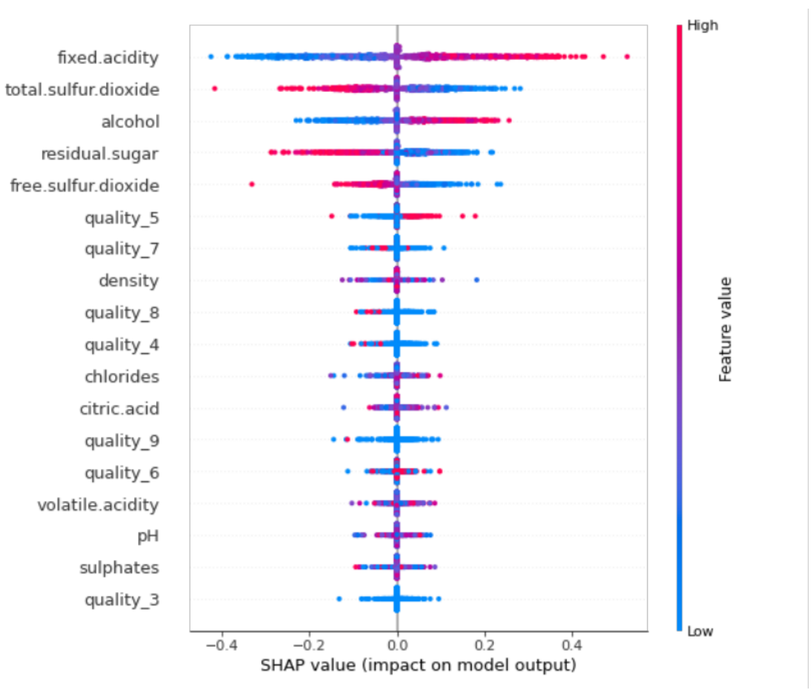


Figure 3: Shapley values per feature value for wine type 2

Entropy

Defintion

Informational entropy or Shannon entropy:

$$H(x) = -\sum_{(k=1)}^{K} p(X=k) log_2(X=k) \tag{1}$$

- We have $K = 3$ classes in this exercise.
- When the algorithm predicts $\frac{1}{3}$ for all three classes, then the entropy is maximized.
- This is because the uniform distribution is the most uncertain.
- A distribution where all the mass is in exatly one outcome, has minimum entropy.

Entropy Example

ID	pred 1	pred 2	pred 3	Entropy
167	0.358059	0.267728	0.374213	1.570201
530	0.400761	0.326552	0.272687	1.567135
473	0.284649	0.424455	0.290896	1.558966
947	0.233179	0.385329	0.381492	1.550319
780	0.419048	0.232465	0.348487	1.545133

(a) Most uncertain

ID	pred 1	pred 2	pred 3	Entropy
866	0.999341	0.000381	0.000278	0.008561
721	0.999180	0.000524	0.000296	0.010363
878	0.999162	0.000546	0.000293	0.010560
831	0.999153	0.000453	0.000394	0.010710
749	0.999103	0.000341	0.000556	0.011229

(b) Least uncertain

Figure 4: Sample outcome of Anomalies

By setting a threshold t we can create a **binary measure of uncertainty**.

$$f(\text{entropy}) = \begin{cases} 1, \text{entropy} \geq t \\ 0, \text{entropy} < t \end{cases}$$