# Week5 tutorial - Anomaly detection:

This week we will work on Insurance fraud detection.

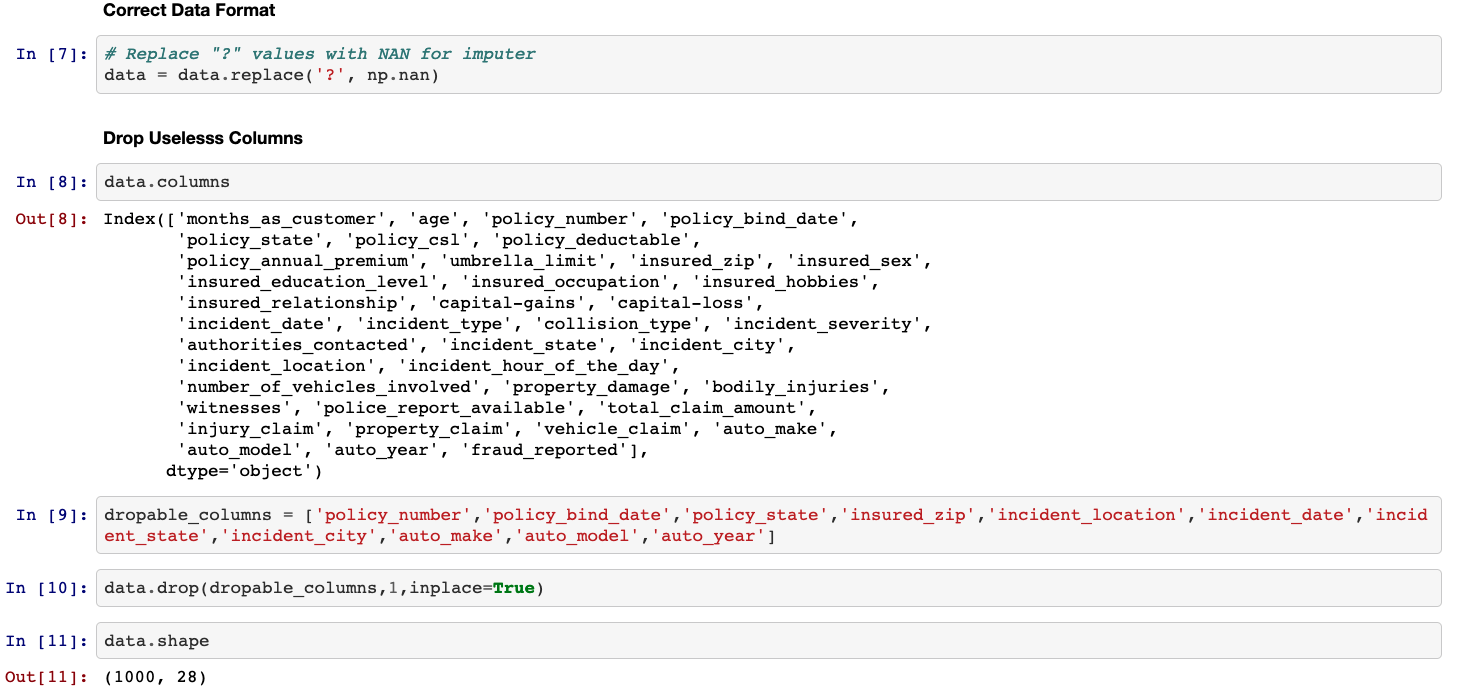
Download the dataset that has been provided (insurance\_claims.csv from https://github.com/moe-ai/SBU\_Data\_Mining).

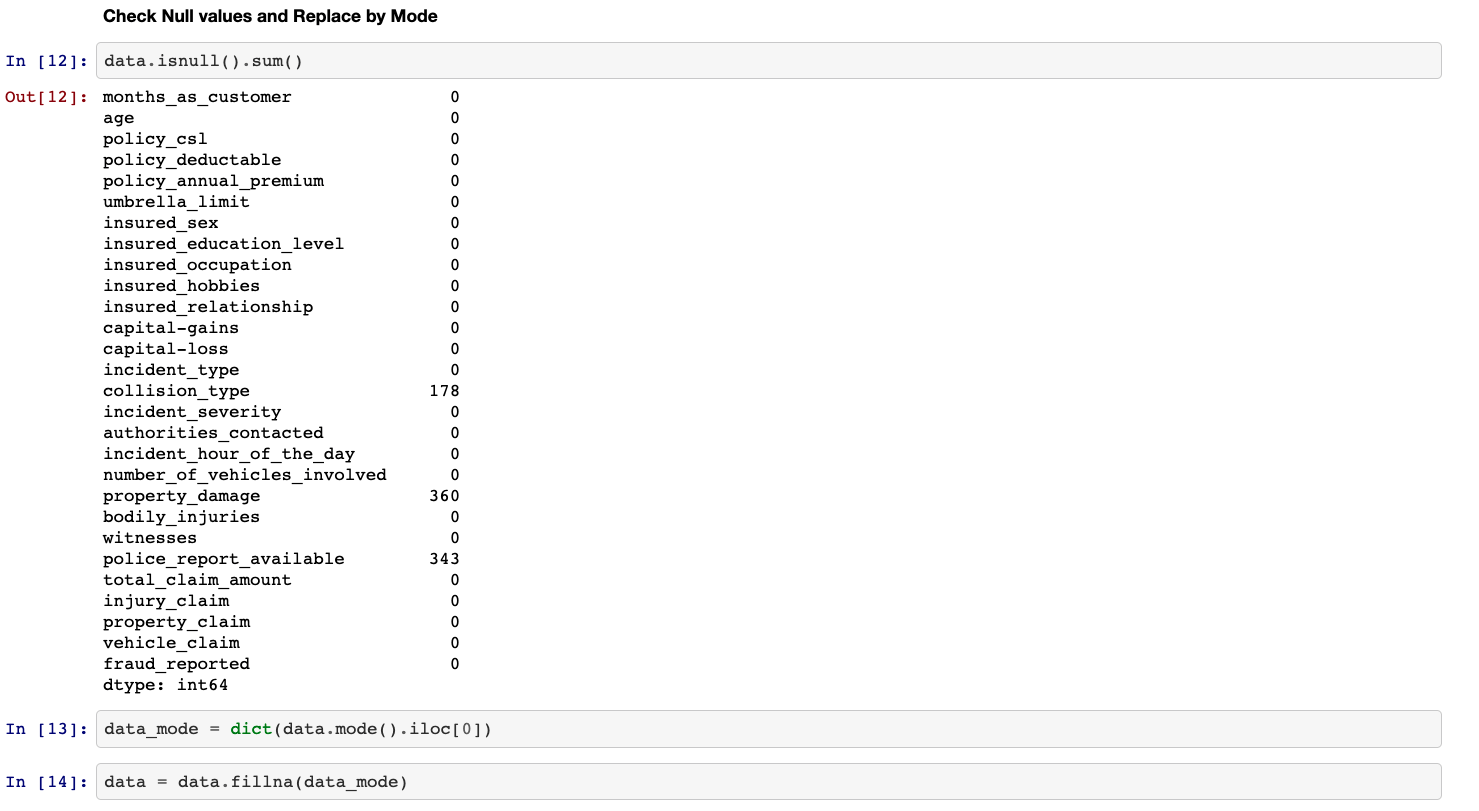
1. First of all, as we have practiced before, we have to load data and do the preprocessing.



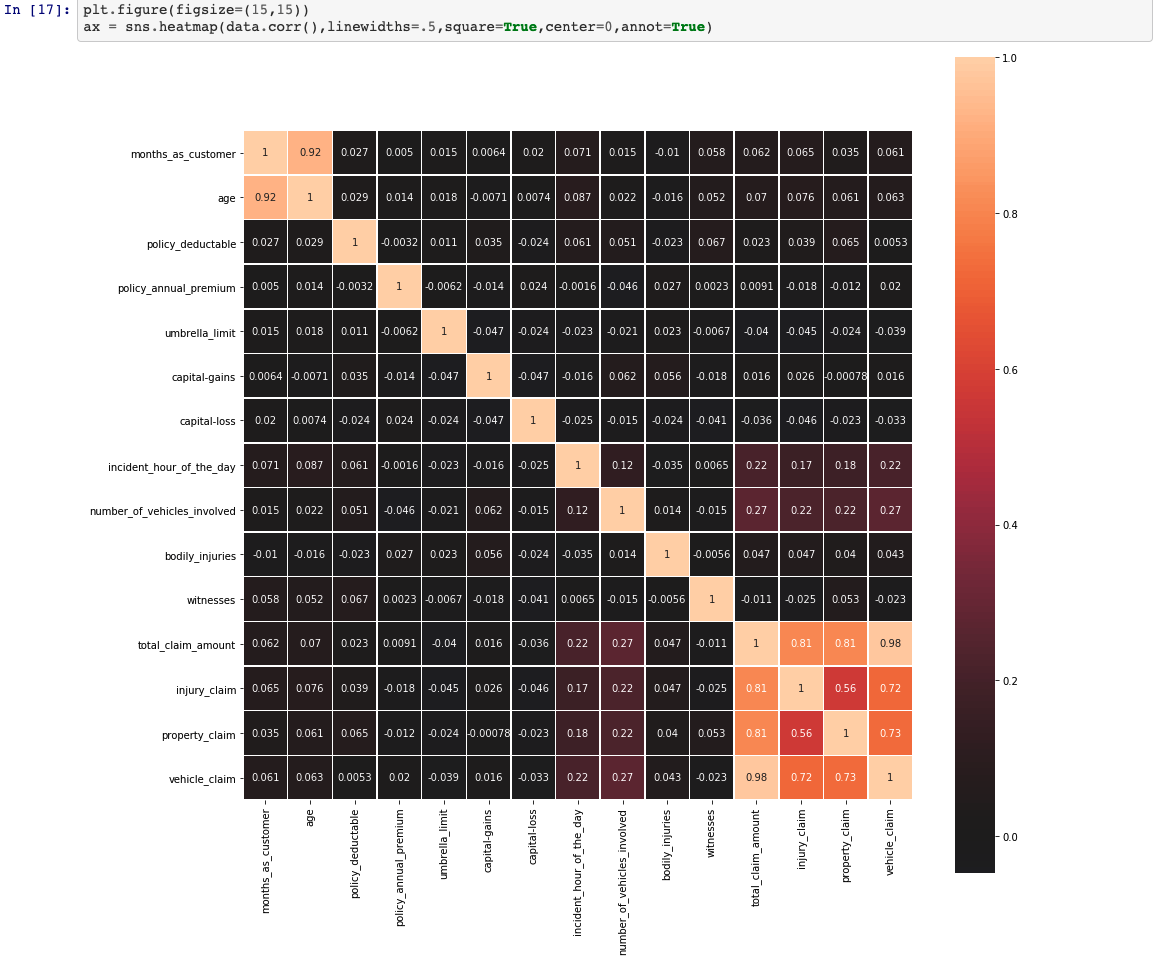
As you can see, we have 39 columns describing the customer and insurance details.

1. Data cleaning:

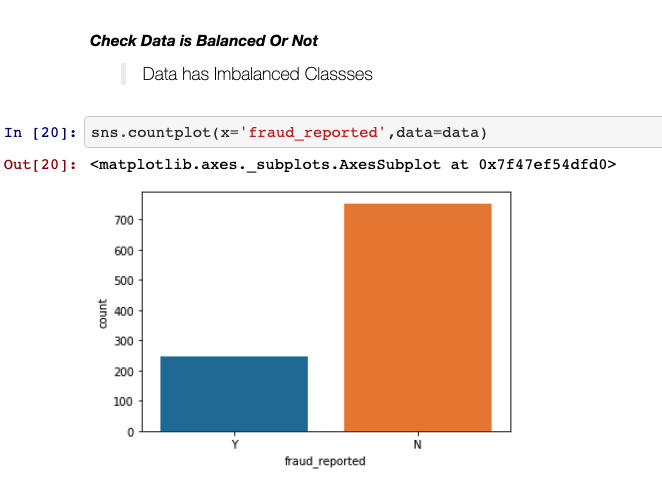




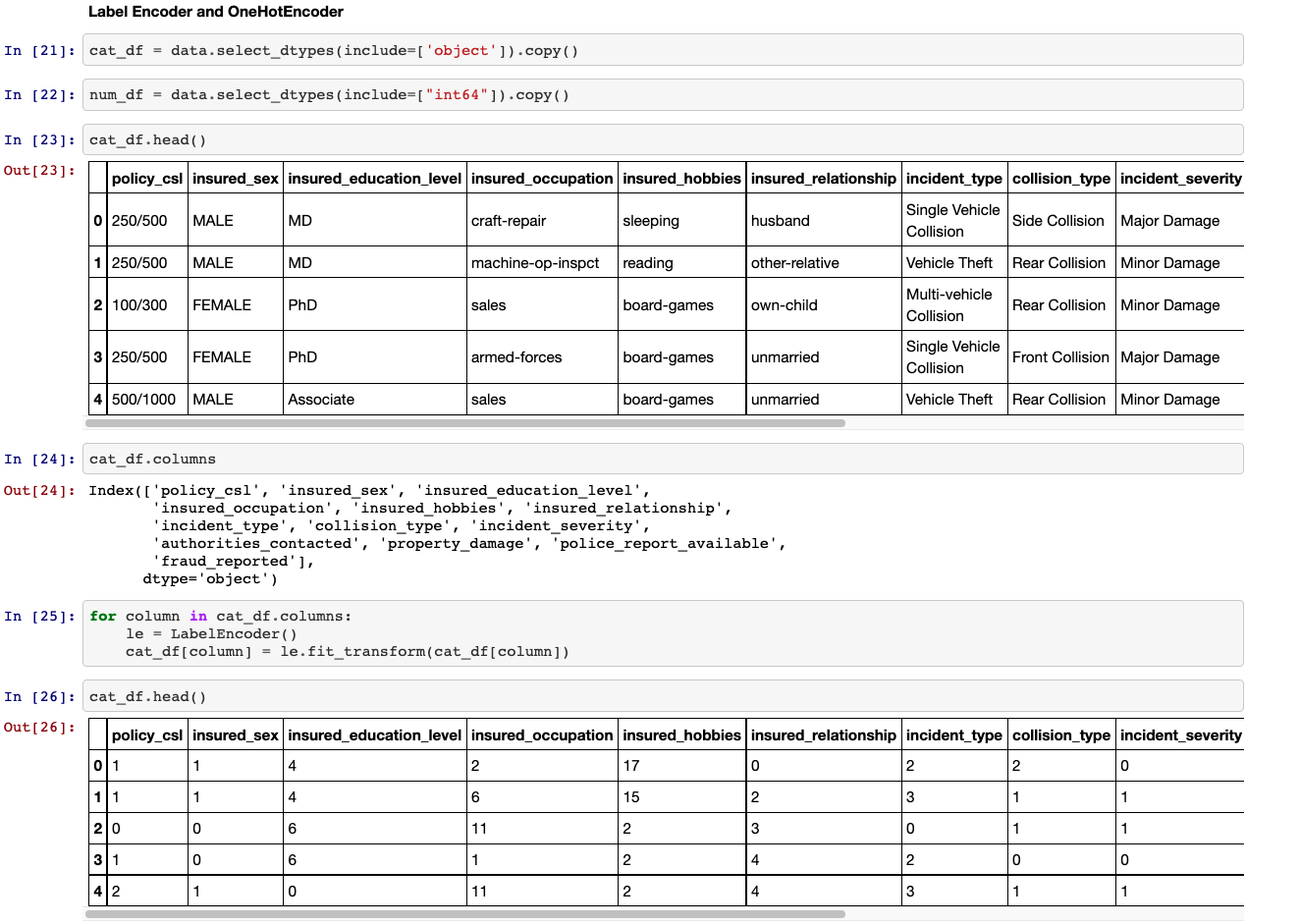
4. Feature’s correlations:

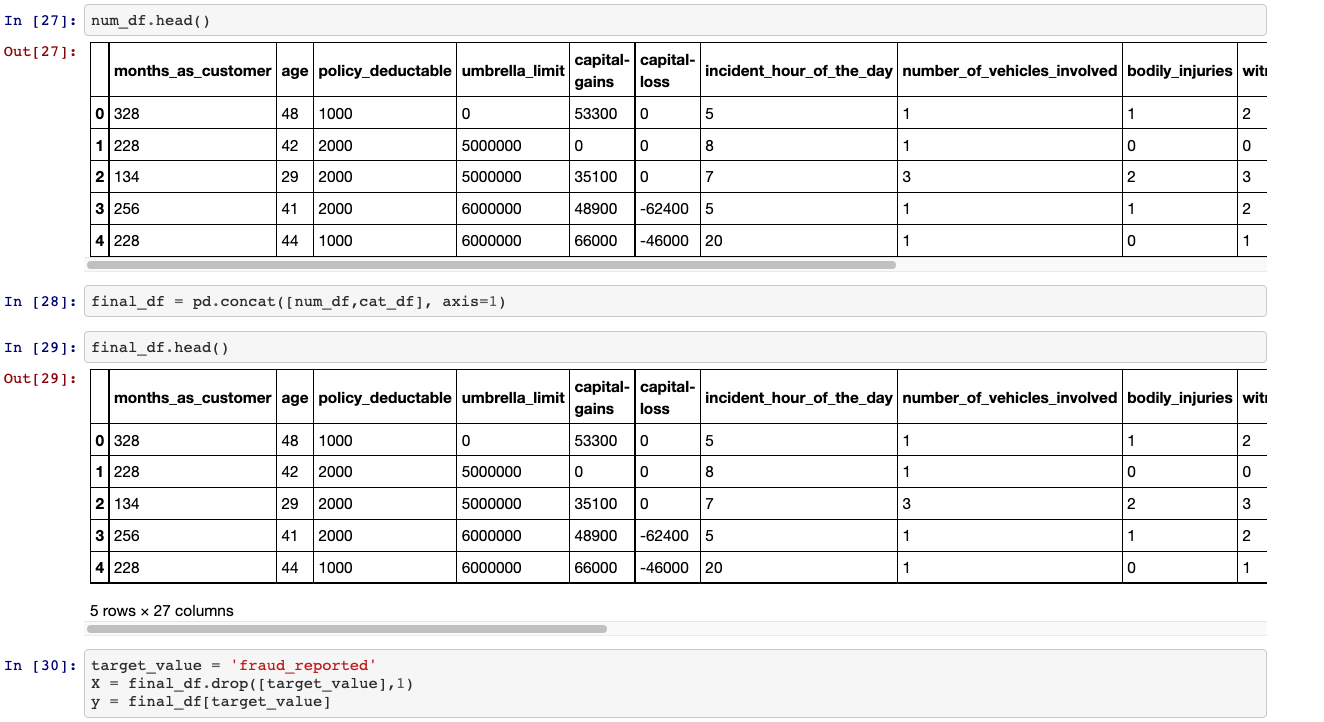


5. Label distributions:



6. Preparing data for training:



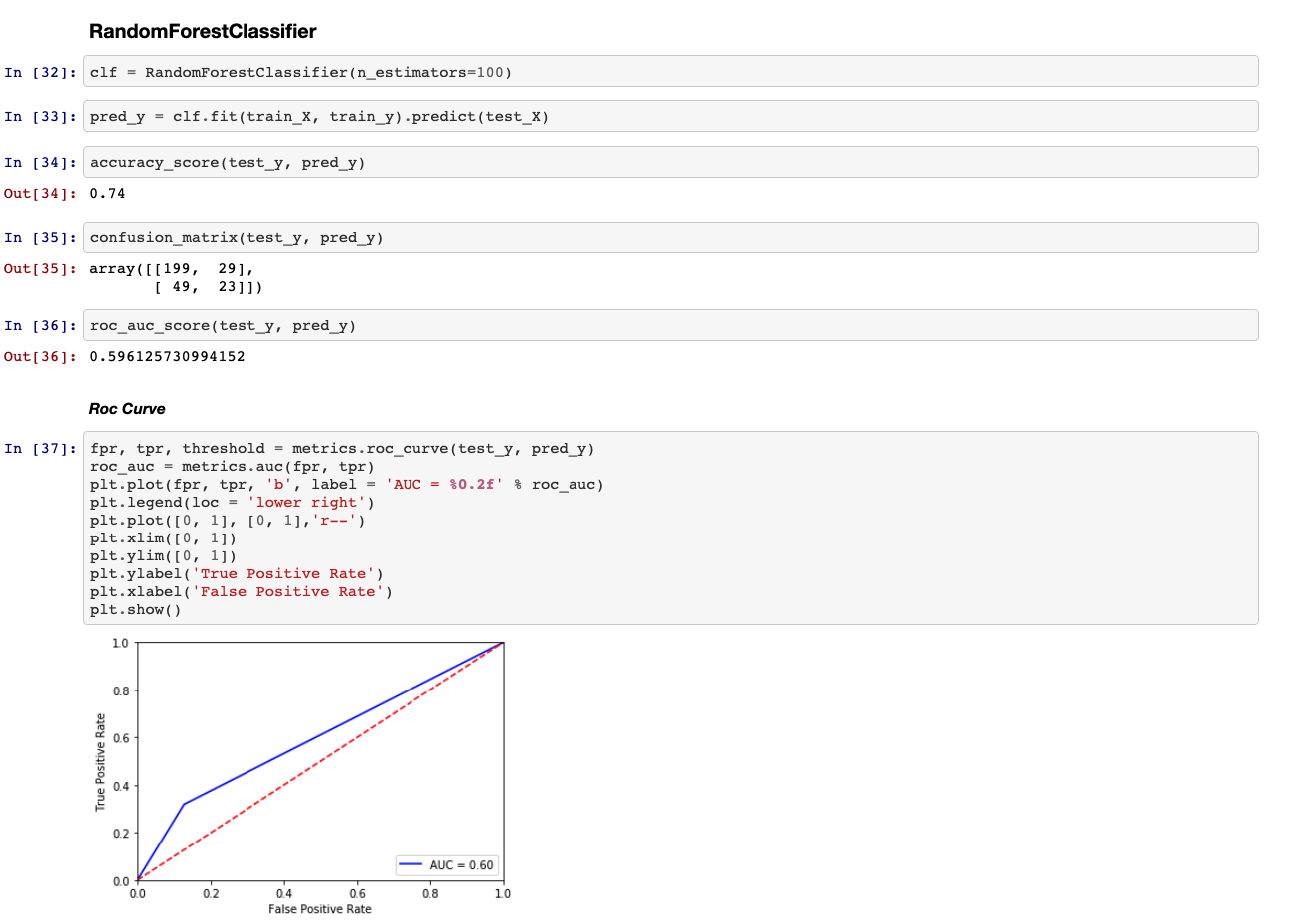


7. Model training:

We apply a random forest classifier. The data has been cleaned and split into train and test datasets in the previous steps.

A receiver operating characteristic curve, or ROC curve, is **a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.** The ROC curve shows **the trade-off between sensitivity (or TPR) and specificity (1 – FPR)**.

<https://www.youtube.com/watch?v=fF0MvCPsEp8&t=122s>



To learn more about ROC curve and AUR, and how to select the best algorithm, please watch the video below (13min to 15min has the main explanation for selection)

<https://www.youtube.com/watch?v=4jRBRDbJemM>

For more examples of fraud detection and anomaly detection check this interesting Github:

<https://github.com/Vicam/Unsupervised_Anomaly_Detection/blob/master/Anomaly%20detection%2C%20different%20methods%20on%20a%20simple%20example.ipynb>