Train In Data

Re-sampling, Cost-sensitive learning and Probability Calibration

ML model outputs and probability

- Logistic Regression returns calibrated probabilities

- Some machine learning models return uncalibrated probabilities
  - Decision trees
  - Naïve Bayes

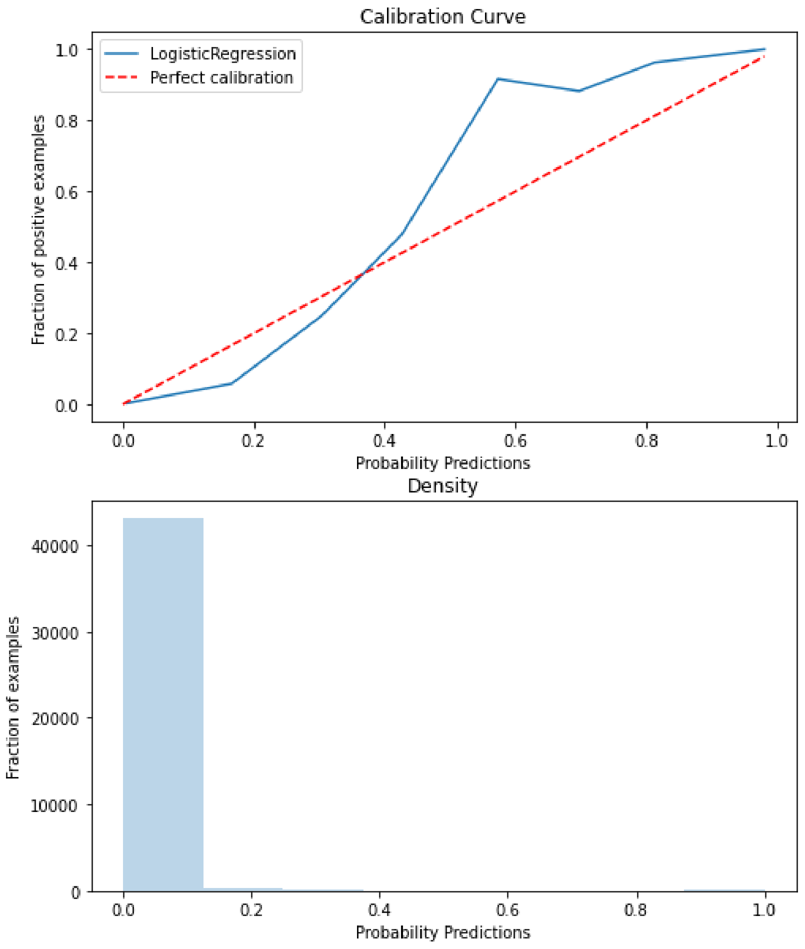
Imbalance data techniques and probabilities

Over-sampling, **under-sampling** and **cost-sensitive learning distort** the relationship between the returned probabilities and the fraction of positive observations.

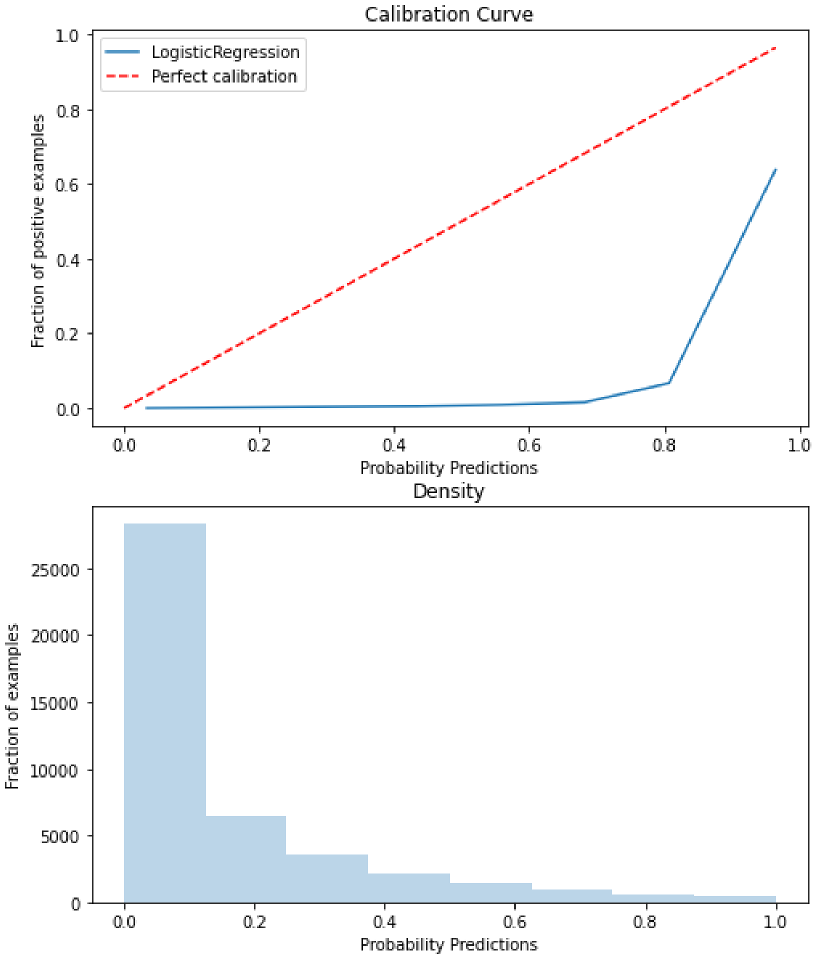
- The first distort the distribution of classes.
- The second modifies the learning function.

Re-sampling – Logistic Regression

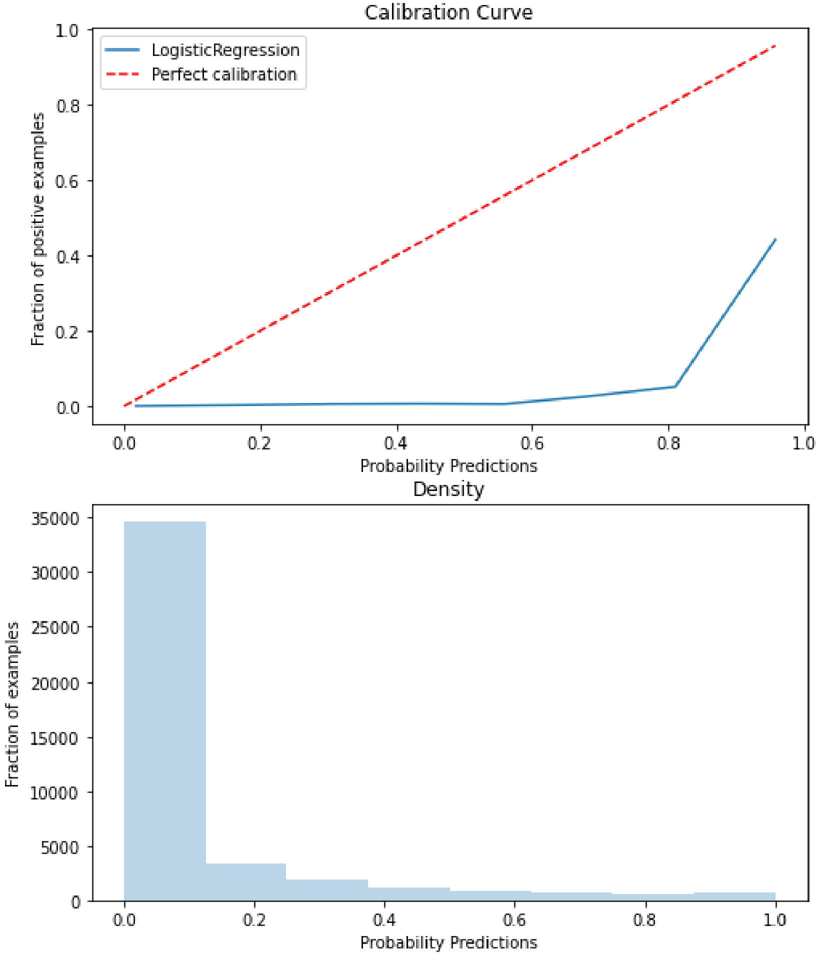
Raw Data



Random under-sampling

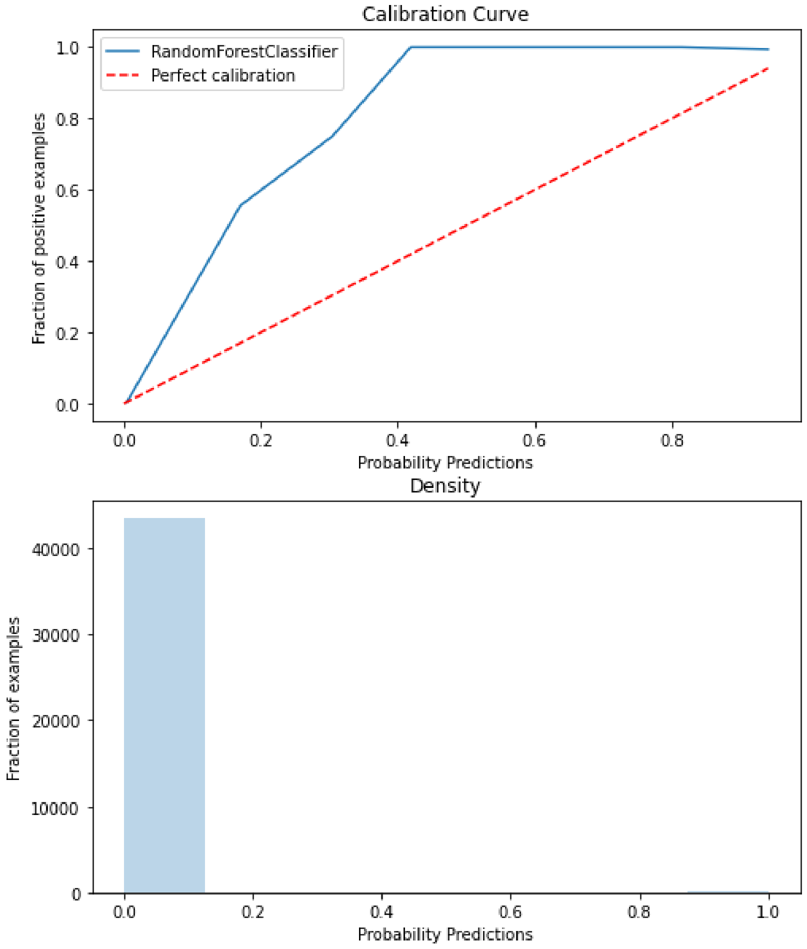


Borderline SMOTE

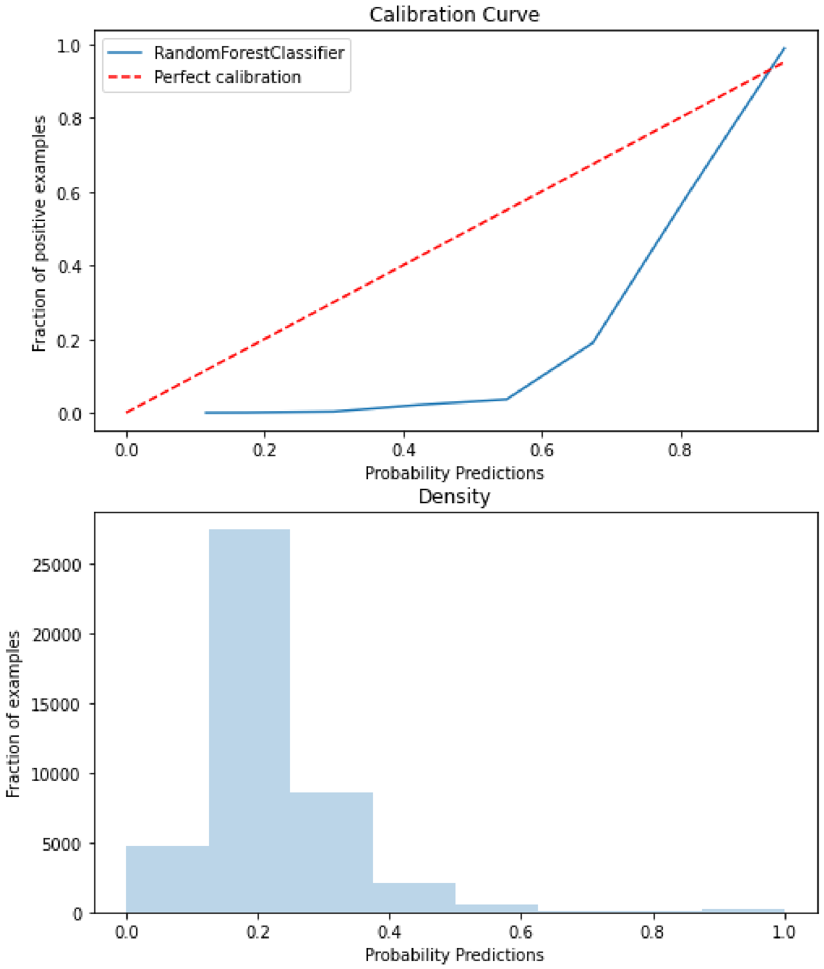


Re-sampling – Random Forest

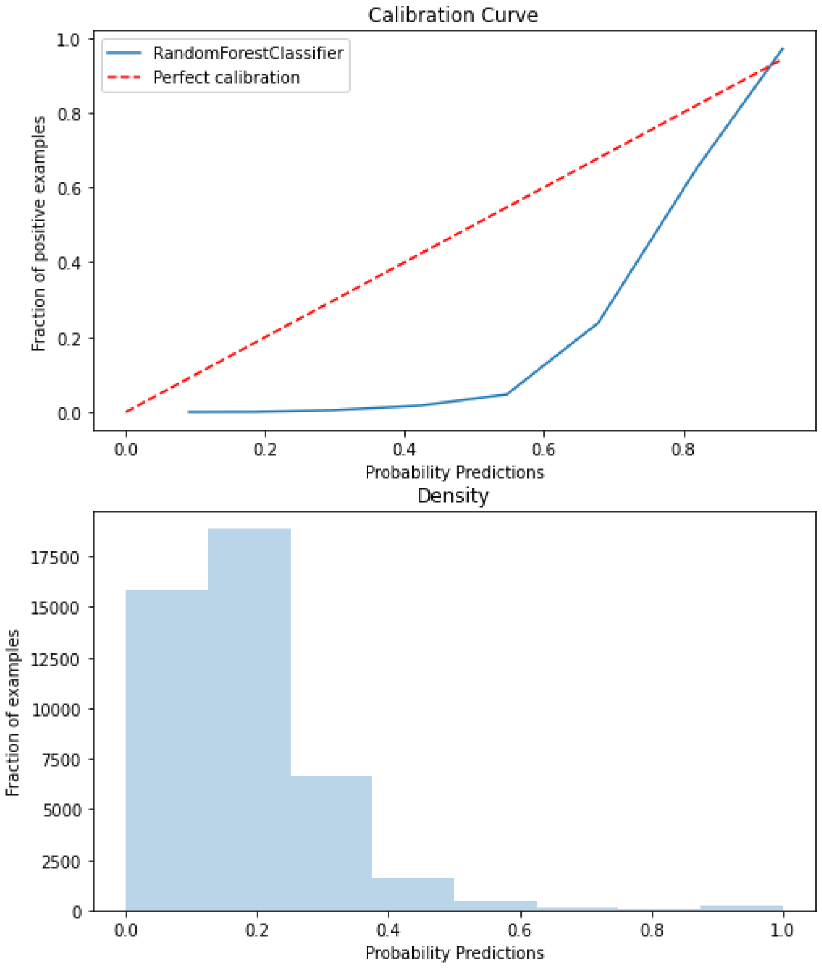
Raw Data



Random under-sampling

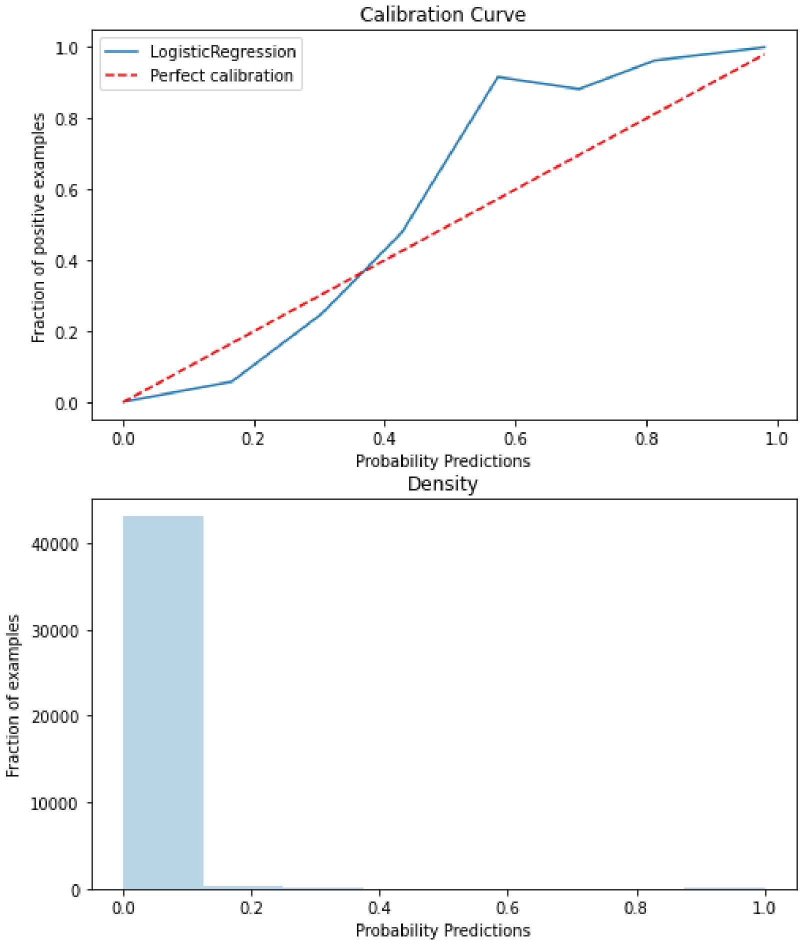


Borderline SMOTE

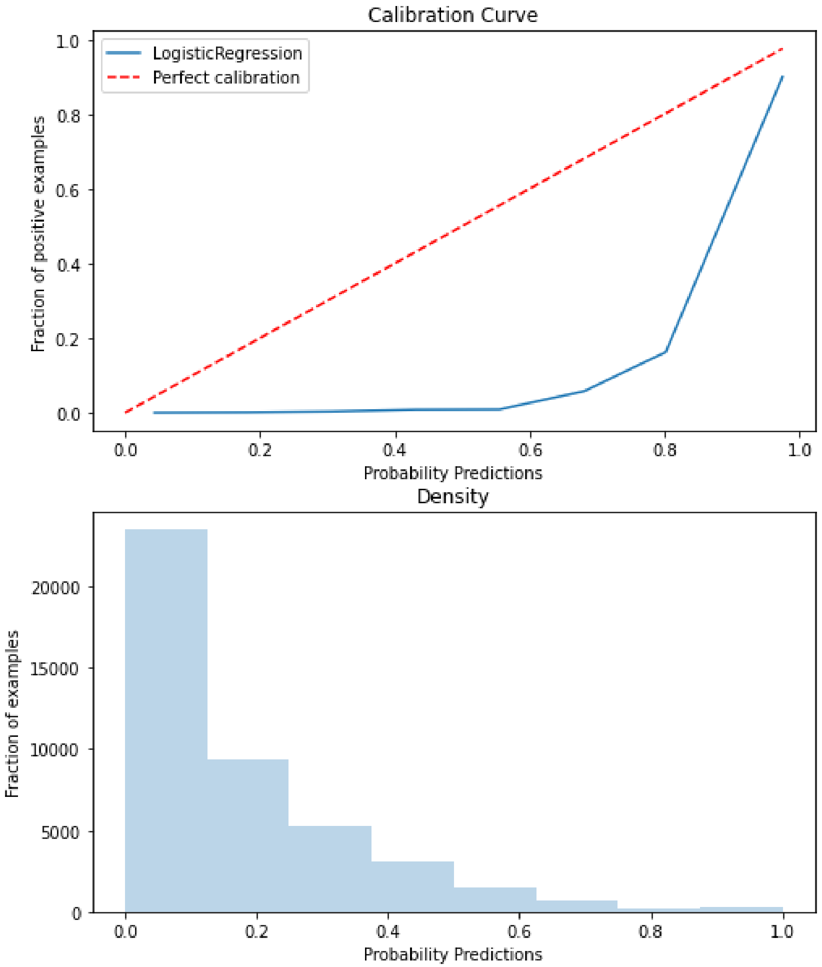


Cost-sensitive – Logistic Regression

Raw Data



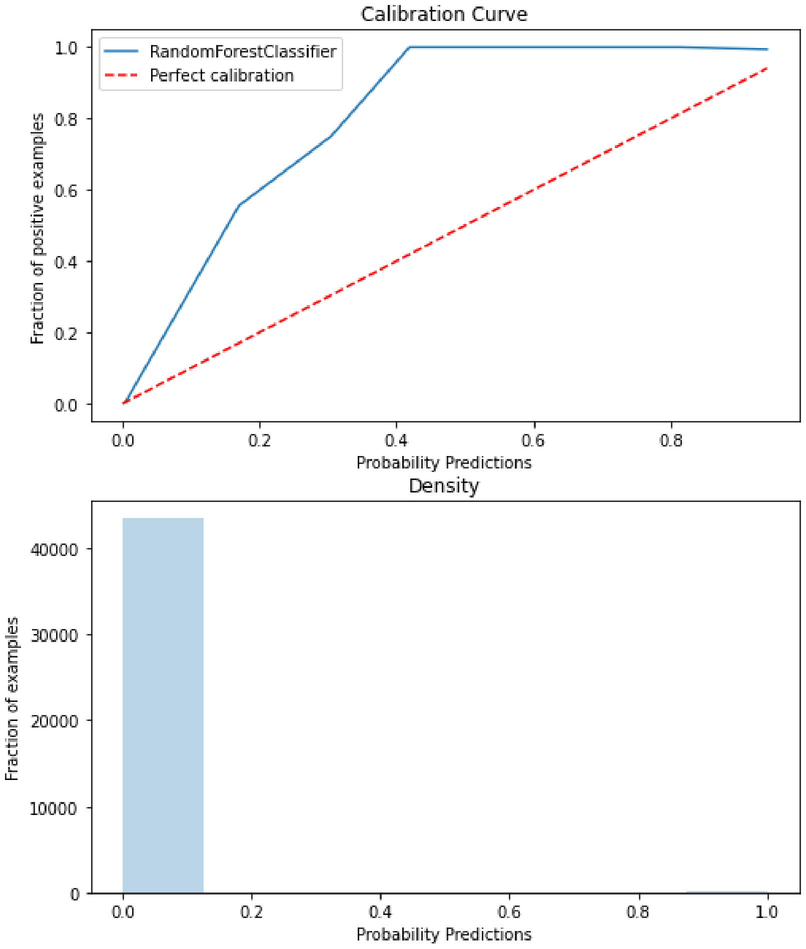
Cost-Sensitive Learning



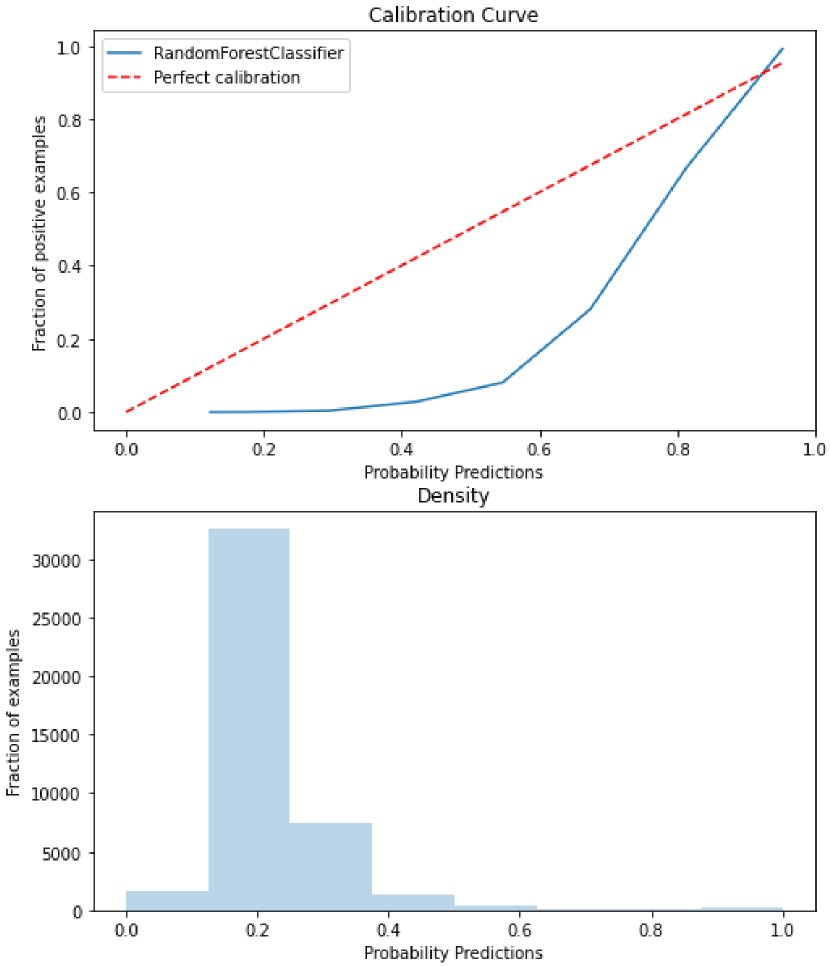
Very similar to Random Under-sampling!!!

Cost-sensitive – Random Forest

Raw Data



Cost-Sensitive Learning



Very similar to Random Under-sampling!!!

Probability as certainty

- Probabilities can be much more informative than labels.

- “The model predicts this claim is fraudulent” vs “The model predicts this claim is 90% likely to be fraudulent”

- To convey likelihood, we need **calibrate** the probabilities after re-sampling or cost-sensitive learning

THANK YOU

www.trainindata.com