Predicting Fraudulent Transactions

Which models predict most correctly with few false positives?

The problem

Company

Particularly credit card processors; potentially, any company with an interest in preventing fraudulent transactions.

Context

Dataset from Kaggle contains 284,807 labeled transactions, of which 492 (0.17%) are labeled fraud. Data have been anonymized by PCA transformation, so little additional human-readable information is available.

Problem statement

Develop a model using machine learning techniques such as logistic regression, to classify and predict fraudulent transactions.

Challenges deep-dive

Challenge 1

Unbalanced classes

Hopefully, fraud is a very small percentage of financial transactions. However, this makes accurately classifying difficult; a model could guess “not fraud” always, and be over 99% correct!

Challenge 2

Context removed

For privacy, data has been transformed to remove all context from the feature columns. Engineering additional features is not possible.

Challenge 3

Measuring accuracy

Related to the class imbalance, accuracy measures must be carefully chosen to focus on correctly identifying positive (fraud) events and reducing false positives.

Solution

Model Stacking

I compared the performance of several models including logistic regression, support vector classification, and the extra-trees decision tree classifier using confusion matrix and ROC/AUC measures. Then I trained a stacked model including all three and k-nearest neighbors. The stacked model produced similar precision with much better recall.

Model Metrics

Measure	Dummy	Logistic	SVC	ExtraTrees	Stack
Accuracy	0.9966	0.9288	0.9300	0.9184	0.9993
Precision	0.0000	0.9126	0.9228	0.9533	0.9472
Recall	0.0000	0.0720	0.0596	0.1046	0.8442
ROC/AUC	0.50	0.98	0.98	0.99	0.98

Precision = True Positive/(True Pos. + False Negative), or correct vs missed fraud ID.
Recall = True Positive/(True Pos. + False Pos.), or correct vs false alarms of fraud.
Goal: maximise precision, then maximise recall, or catch all frauds with few false alarms.

Impact

Vastly fewer false positives using stacked model

The stacked model correctly identified 466 of 492 fraud cases in the dataset, for a precision of about 95%.

Only the ExtraTrees model performed a tiny bit better, correct on 469 cases. Even the least-precise individual model had a precision of about 91%.

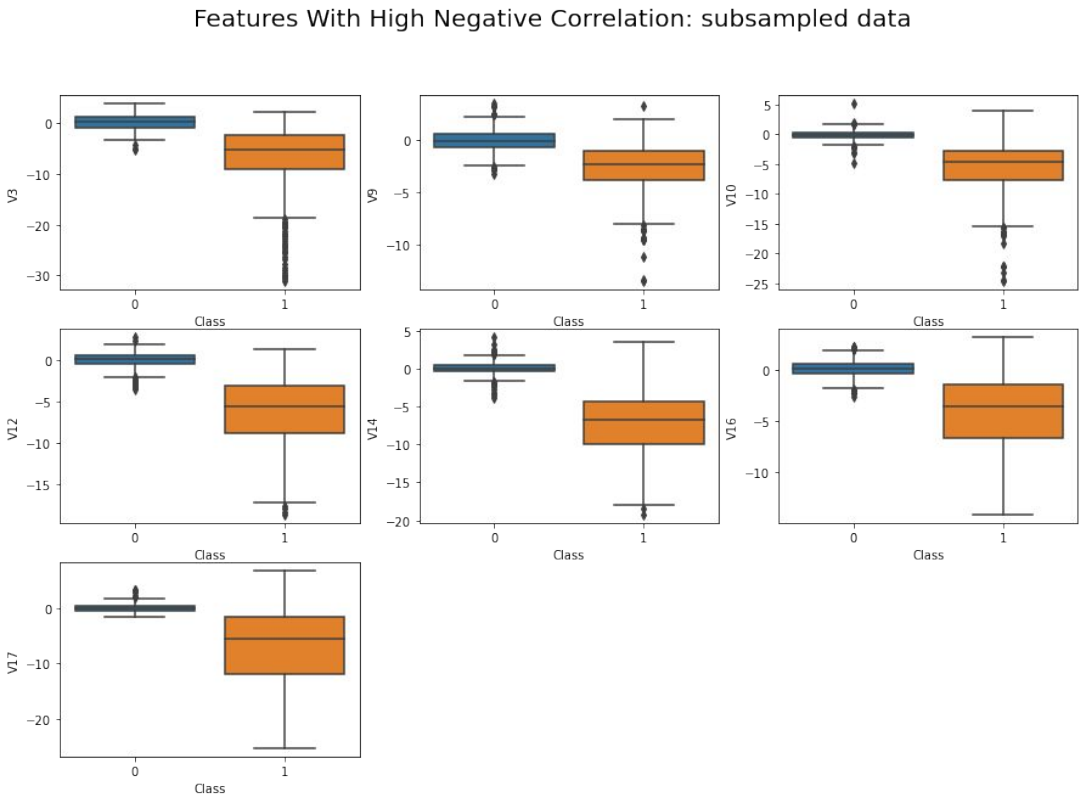
However, all the individual models had false positives in the thousands and recall rates of 10% or less.

The stacked model had only 86 false positives, for a recall rate of 84%.

This is not only 8x better recall, but up to 7,000 fewer confirmation calls/tickets!

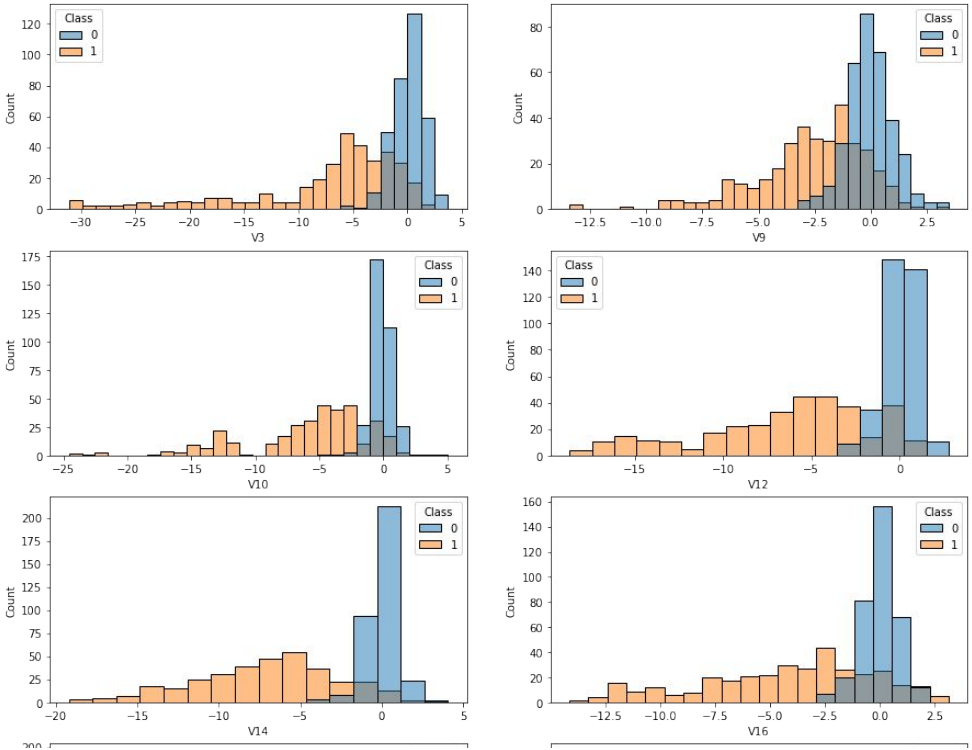
Interesting Questions

What are the highly correlated features?



I noticed some interesting-looking patterns in features with strong correlations. If I had access to production data, I would want to know which features produced strong correlation. Can we figure out why the fraud (orange) cases have so much wider distribution and yet center so much lower/higher than the non-fraud? Can we use this answer to inform our model selection and training?

Can we make classification easier?



This is a view of the distribution of some of the strongly-correlated features, colored by class. Fraud class is orange. You might notice that while there is overlap, there are many columns in the fraud class that are distinct from the non-fraud class. Can we encourage our model to focus on these features? Again, if I had access to feature names, would there be information here that might help human reviewers with the more difficult cases?

Thank you!