

# Why Tree-Based Models Are a Popular Choice for Fraud Prevention

Ekata | 4.30.2018 | [in](#) [twitter](#) [facebook](#) [link](#)

Fraud is just one of an infinite number of problems machine learning can help solve. But choosing the right model for a specific use case like credit card fraud, promo abuse, or new account fraud can be difficult. And businesses must often build multiple machine learning (ML) models to battle fraud effectively.

Today's post highlights tree-based models and why they are a popular choice when it comes to preventing fraud.

## A Popular Choice for Fraud Prevention

Tree-based models are a popular choice for fraud prevention compared to other models for many reasons. One reason is that tree-based models are typically easier to work with than other models like logistic regression because trees require less feature engineering. Tree-based models are also good at handling categories of variables (e.g. match/no match, true/false, etc.) and they generally have higher accuracy when it comes to predictions compared to regression models.

Another advantage of using tree-based models is that they are good at picking up relationships. For example, if you build a model for selling a house, a regression model won't pick up that the higher number of bedrooms, the greater the total square feet, but a tree will pick up that some of the variables are related. Regression models view each variable as independent and do not recognize relationships like tree-based models.

While there are a number of different types of tree-based models, two of the most popular tree-based models are Random Forest and Gradient Boosted Tree. Both are ensemble models which typically provide greater accuracy and broader coverage than single decision trees. This is because ensemble models involve building multiple models instead of a single model. The models are combined into one generalized model that is less impacted by training data outliers.

## When Would You Use a Regression Model?

So why would anyone use a regression model, specifically logistic regression, for fraud prevention or any other use case? – Explainability. With logistic regression, you can explain how one variable is more powerful than another, and you can't easily do that with a tree-based model. In addition, logistic regression works well with smaller datasets or datasets where the signal could be easily lost in the noise. These attributes make logistic regression a great starting model while you work to build up your training data.

## Confidence Score Uses a Tree-Based Model

[Confidence Score](#), one of the features of our [Identity Check](#) product, uses a Random Forest model to calculate a predictive risk score ranging from 0 to 500. In most cases, the higher the score, the riskier the transaction. One of the ways Confidence Score can be accessed is as an attribute in the [Identity Check API](#).

Tree-based models work well with Identity Check because they handle combined signals well, tolerate missing or unnormalized data, and require less feature engineering than regression models. Relationships between attributes such as email, phone, person, address, and IP must be taken into account when assessing most types of fraud. And tree-based models are good at detecting relationships between variables.

## The Model is Just One of Many Pieces

No one ML model is going to solve all your fraud problems. You also need the [right data](#) to feed each model, a data scientist to run experiments and fine tune each model, and a manual review process to verify that transactions flagged as fraud are actually fraud. The model is just one of many critical pieces of an effective fraud prevention arsenal.

To learn more about how we use machine learning in our products, check out our [Ekata Machine Learning Guide](#).