Executive Summary

PayPal is a global company operating a worldwide online payments system. The company’s innovative open digital platform gives its 218 million active account holders in 202 markets across 25 currencies the confidence to connect and transact online.

Fraud prevention is an important area of investment for PayPal. The company has successfully used machine learning and deployed robust fraud prevention models for more than 10 years. However, fraudsters are constantly changing their patterns and uncovering new ways to take advantage of the system. As a result, PayPal must continuously find ways to improve fraud detection accuracy and decrease fraud detection time.

Using H2O Driverless AI, the PayPal team was able to find significant new modelling features which dramatically increased model accuracy by almost 6% in a single test. For a team with over 10 years of feature engineering experience, this was an amazing result. PayPal plans to continue to use Driverless AI in innovative ways to prevent fraudulent activities.

The Challenge

To protect its consumers from fraud, PayPal offers an extensive purchase protection guarantee for buyers, promising to reimburse them for the full purchase price plus any original shipping costs if they fail to receive the item they ordered. A similar protection guarantee extends to merchants through the seller protection program, which helps guard sellers against loss due to claims and chargebacks. Unfortunately, with such a high transaction volume, PayPal has experienced fraud from buyers and sellers colluding to defraud its protection programs.

PayPal’s approach to detecting fraud includes using teams of data scientists, financial analysts, and external intelligence agencies to learn how fraud perpetrators think, what drives them, and techniques they may attempt to use to exploit PayPal’s payment system. These teams collaborate to build robust models aimed at predicting and preventing unlawful activity.

The Solution

The PayPal data science team has worked with H2O.ai, the open source leader in AI, for a number of years, using machine learning technology and statistical models to detect fraud patterns. To stay one step ahead of fraud perpetrators, PayPal challenged itself to look at the problem differently, examining not only individual buyers’ and sellers’ behaviors but also considering activities that seemed to indicate an association with larger, interconnected networks. For example, are the suspect buyers and sellers sharing assets? Do they share the same IP address? Are they listed under the same shipping address?
To better understand these new networks of data, PayPal's team implemented a graph database (neo4j) and used node2vec, an algorithmic framework for learning continuous feature representations for nodes in networks. Node2vec uses the notion of node network neighborhoods and explores how nodes can be organized based on the communities they belong to or based on the nodes' structural roles in the network.

In PayPal's case, a node can be the account number or the IP address of a buyer or a seller. Once a bad account is identified based on payment transaction data, other bad accounts sharing the same network structure can be located. Even with substantial feature engineering and model training on this new data, the data science team was not satisfied with the results.

Having worked with H2O.ai for years, the PayPal team turned to H2O Driverless AI to see if the platform's automatic feature engineering could help build a more robust model. PayPal combined feature representation from the graph network structure with their expert-engineered features and then applied H2O Driverless AI to the merged feature set. Driverless AI automatically engineered additional features and models, greatly improving model performance. Using a three months subset of payment transactions data, PayPal looked specifically for collusion fraud. The size of the graph was about 1.5 billion edges and 1/2 million nodes and the number of features is around 400 to 600, with Driverless AI doing the model training and automated featuring work.

<table>
<thead>
<tr>
<th>DATA</th>
<th>ENVIRONMENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training Data:</td>
<td>Driverless AI: Feature engineering and model</td>
</tr>
<tr>
<td></td>
<td>training.</td>
</tr>
<tr>
<td>Test Data:</td>
<td>Spark: Data preparation and pre-processing</td>
</tr>
<tr>
<td>Number of Features:</td>
<td>400-600</td>
</tr>
<tr>
<td>Hardware:</td>
<td>IBM Power 8 GPU server</td>
</tr>
</tbody>
</table>

**The Results**

The top 5 features extracted by H2O Driverless AI bested 10 years’ worth of expert engineered features. At the same time, H2O Driverless AI increased model accuracy increased from .89 to .947. In addition, running on an IBM Power GPU-based server allowed the team to train the model six times faster when compared to a CPU environment.

**Next Steps with H2O Driverless AI**

Applying machine learning directly to graphs with H2O Driverless AI opens exciting new possibilities for PayPal. One of the company’s immediate goals is to evaluate Driverless AI directly with raw data by plugging it into the data stream using time series functionality to eliminate manual feature engineering on new data.

**Ready to see it in action?**

Driverless AI is helping companies achieve real business benefits using AI. To learn more about H2O.ai and to request a live demo of Driverless AI, visit us at www.h2o.ai.