QuadPay Data Scientist Assignment

As a Data Scientist at QuadPay, you'll help solve interesting problems on a daily basis, including fraud prevention and building real-time credit-decisioning models. The models you build will power a platform that is processing millions of transactions every month across the largest e-commerce brands in the USA.

For this assignment, you'll be supplied with an anonymized dataset of historical orders and their repayment status.

You'll analyze the data and look for any insights that may help us design a classifier that can be used to approve future customers to pay in installments. These give you a chance to demonstrate your engineering, data science and communication skills.

We expect you to treat the assignment like a day-in-the-life at QuadPay. You're free to use the tools you are familiar with and whichever resources that will help you get the job done.

Data Challenge

QuadPay is a payment gateway that lets consumers split purchases into 4 interest free installments, every two weeks. The first 25% is taken when the order is received, and the remaining 3 installments of 25% are automatically taken every 14 days. We help customers manage their cash-flow and we help merchants increase conversion rates and average order values.

This assignment is designed to help you become familiar with our problem domain and start to think about which scenarios we should anticipate going forward. It gives us an opportunity to evaluate how you approach complex problems.

Training data

Orders.csv contains an anonymized set of customer orders, labelled with details about which installments the customer paid. It has the following columns:

- order_id : String
- customer_id : String
- merchant_id : String
- order_amount : Decimal
- checkout_started_at : Datetime
- credit_decision_started_at : Datetime
- approved_for_installments: Boolean
- customer_credit_score: Integer

- customer_age : Integer
- customer_billing_zip : String
- customer_shipping_zip : String
- paid_installment_1 : Boolean
- paid_installment_2 : Boolean
- paid_installment_3 : Boolean
- paid_installment_4 : Boolean

Note that customers may have multiple orders in this dataset.

Objective

An order is considered in defaulted if any of the installments have not been paid. Our aim is to keep approval rates high (allow as many customers to transact as possible) while reducing the total defaulted payments.

Using your tool of choice (eg Python, R, Jupyter notebooks, Matlab, Tensorflow, etc), ingest the data and look for features that could be good candidates for an InstallmentApproval model which we could use to determine whether a customer should be allowed to make a purchase.

With such a small training dataset, we're not expecting the world's most accurate analysis! We're interested in how you break down a problem, appropriately communicate your findings, and what you identify as good next steps for future analysis.

Please submit a code repository with instructions for how to execute your mini-pipeline that generates the results and charts that you used in your analysis.

Questions to explore:

- Which features show strong correlation with a customer's likelihood of paying back installments?
- Which features should be discarded? Why?
- What surprised you about the results/trends observed in the data?
- What additional data would you like to see that might help build a better installment-approval classifier?
- What would be your next steps to train/build a model that we could use to make real time customer approval decisions?

Submission

Please send us your final code repository. We'll schedule a follow-up call to discuss your findings.

Good luck and have fun.