Case Study

# The Task

For a dataset of clients, predict their propensity to default on loans using a model of your choice. This task has two distinct parts:

1. Build a model to estimate that target, AND
2. Describe your process

When describing your process, pay attention to the following:

1. Describe any data manipulation you performed
2. Describe the algorithm you chose and the results
3. Describe the performances of the model

Please make sure to submit both **the code file** (Python script if possible, or otherwise in whatever language you prefer) and **the text file** containing your process description.

For any questions, please feel free to contact [einat.aviv@bluevine.com](mailto:einat.aviv@bluevine.com).

# About the Dataset

The dataset is based on real BlueVine clients. It includes 17 features and a target variable: *tag\_in\_six\_months*. The dataset is split into two time periods and includes clients from these two different periods according to their seniority (the same client might appear twice, once in every period).

Fields:

* client\_id: unique Client identifier.
* pit: Point in time. The timestamp associated with the row.
* period: We have two points in time which represent the seniority of the client since the first funding date. First\_funded+180 AND First funded+120. You should consider developing two different models, one for each period.
* recent\_successful\_repayments: number of successful repayments made by the client up to the row’s point in time.
* future\_bad\_probability: The results of the existing version of this model, which is based on all the features in this dataset. Your model is intended to be the “challenger” for this “champion” model.
* credit\_score: A client scoring metric from a credit report provided to BlueVine by an external data provider.
* available\_credit: client’s available credit according to the credit report referenced in the credit\_score field.
* balance: Client’s bank balance at the row’s point in time.
* sum\_failed\_repayments: dollar sum of the Client’s failed repayments (repayments that was supposed to have made but actually didn’t make) at the row’s point in time.
* max\_failed\_repayments: the dollar amount of the Client’s largest failed repayment.
* external\_score: another credit scoring metric.
* credit\_inquiries\_count: Amount of searches performed on the client according to the credit report referenced in the credit\_score field.
* credit\_open\_balance: sum of balances in open accounts according to the credit report referenced in the credit\_score field.
* years\_on\_file: An indication of how many years the client has been on file in credit report referenced in the credit\_score field.
* num\_successful\_repayments: number of successful repayment made by the Client according to the credit report referenced in the credit\_score field.
* missing\_credit\_report: binary indicator if the credit report is missing or not.
* client\_industry\_unkown: binary indicator if the Client’s industry is known or not.
* tag\_in\_six\_months: This is the target variable. Client’s tags after six months. Can be one of the following: bad (=default) OR good (=not default).