Phase 3 Draft proposal

Current state of delivery after Phase 1 of ML 2

At present, 4 clients consisting of 5 setups have gone into production where break and comments are predicted for the users using machine learning model. Due to the nature of reconciliation and the vast amount of data variability possible, all setups have separate ML prediction models.

Thought process before going into Phase 2 of ML 2

After a series of discussions, we have come to the conclusion that there is a possibility of developing a base model which deliver a base accuracy, irrespective of whichever data is given to the model from any client-setup combination.

This hypothesis is based on the assumption that reconciliation as a whole is matching of similar transactions based on inherent variable properties. Once we look at an exhaustive set of all representations of data the model can get, we can safely identify variable properties controlling reconciliation process.

During phase 1, we did not have the entire representation of data at our disposal, therefore all the variable properties were not identified. Rather, the variable properties were identified in on a setup level. Due to more experience with dealing of recon data now, our team is planning to handle 10 times the client base handled earlier in a shorter time frame.

Proposed state of delivery after completion of Phase 2 of ML 2

At the end of this phase, EpowerX would have delivered a minimal viable model which would give the base accuracy to any client-setup combination data given to the model. For achieving a higher accuracies for each setup, specific setup models would have to be developed based on client interactions with the EpowerX team.

Solution not proposed in Phase 2 of ML 2

Since the aim of this phase is to develop a minimal viable model, we do not propose to deliver accuracy any higher than what is proposed below. This model would also not be productionized to be delivered to the end user, so scalable and robust data pipeline is not to be developed during this phase. To execute the end result of the model developed during this phase, Viteos resources would have to work with EpowerX and understand how to execute codebase without API endpoints.

Cost:

Hourly rate per resource : $29

Number of resources : 3

Total number of hours per week per resource : 8 hours/day X 5 days = 40 hours

Total cost for project = $29 X 40 X 4 X 4 X 3 = $55,680

Duration:

Total duration : 4 months

Total clients to be analysed : 50

Base accuracy to be delivered :

Break prediction

UMR

One to One

One to Many

Many to One

Many to Many

UMB

One to One

One to Many

Many to One

Many to Many

Close

Comment

Break up of various tasks for doing the project :

1. Client selection + data sanity check
   1. Explanation: Viteos and EpowerX resources to work together to select all clients with labelled data which can be incorporated into machine learning model. Selection of clients will be based on data sanity check of the clients. Data sanity check will be scalable to analyse client population faster.
2. Variable analysis
   1. Explanation: Based on knowledge from Phase 1 as well as research and business interaction, key variables will be identified around which model can be built.
3. Variable cleaning and selection
   1. Explanation: Standard pre-processing operations in ML model building.
4. Research client bucketing approach
   1. Explanation : The final model will have different models for different kind of buckets. There can be various characteristics to identify buckets like client data size in a given day (small data size < 1000, medium data size from 1000 to 5000 and large data size for above 5000 records, or asset class (futures, dividends, interest, derivatives etc).
5. Actual client bucketing for sample (smaller) client population
   1. Explanation : Once buckets are made, a smaller sample client population (For example, 5 clients out of 50) will be bucketed to test the hypothesis. The hypothesis would be to check if
6. Iterate over clients to group into different buckets and check for best strategy to put a client in a specific bucket
   1. Explanation : Once hypothesis has been tested, iterative bucketing of clients to see which bucket strategy makes the best sense for bucket characteristics. This will further enhance bucketing methodology for future test clients in pipeline.
7. Present results to business
   1. Explanation : Explain technical workflow of bucketing approach to business. Take client suggestions and revamp code to incorporate changes if deemed necessary.
8. Prepare a minimal viable model delivering basic accuracy to all clients
   1. Explanation : The goal of this phase is not to get to highest achievable accuracy. The goal of the phase is to get to a point where we can safely test that a base model based on bucketing approach can be developed which can predict class categories up to 70% accuracy for all classes in break prediction and comment categories. Therefore, once all the above steps have been completed, the final code for minimal viable model will be ready.
9. Test model on test clients
   1. Explanation : To check if bucketing can be applied to a new client for onboarding, we will test the minimal viable model on a separate client base (which come under the test category) and check if achieve required accuracy of 70%.
10. Iterate steps 2 – 9
    1. Explanation : Since machine learning model development is an iterative process, we will iterate steps 1 to 9 till all test clients have the base accuracy levels
11. Approval process
    1. Explanation : Deliver accuracy achieved for all clients to business and get approval for closure of project
12. Document and deliver
    1. Explanation : For final delivery, we will deliver final code with documentation and knowledge transfer.