In Depth Data Analysis proposal

Part I : Introduction

Section A. 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 models. Due to the nature of reconciliation and the vast amount of data variability possible, all setups have separate ML prediction models.

Section B. Thought process before going into Data Analysis phase

After a series of discussions, we have come to the conclusion that there is a possibility of developing a base model which delivers 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 on a setup level. Due to more experience with dealing of recon data now, our team is planning to handle 5 times the client base handled earlier in a time frame of 5 months.

Section C. Proposed state of delivery after completion of Data Analysis phase

At the end of this phase, EpowerX would have analysed the data for 20 representative clients. The deliverables would be :

1. A confirmation about whether a minimum viable model can be developed which can deliver a base accuracy for any kind of recon data inputted into the model.
2. Appropriate bucketing strategy would be developed basis which minimum viable models for each bucket would be developed in future.
3. As proof of working of this strategy, one single model for only one bucket would be given, along with the accuracy expectations from the minimum viable model.

Section D. Future scope after Data Analysis phase

Once data analysis phase wraps up and EpowerX data scientists have tested the hypothesis that minimum viable model approach is feasible, both parties can extend the contract for actual development and delivery of a minimal viable model. This model would predict recon results with a definitive base accuracy for 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.

Section E. Solution not proposed in Data Analysis phase

Since the aim of this phase is to analyse data, we do not propose to deliver the minimum viable model at the end of this phase. This phase is supposed to be a quick experimentation phase to test the hypothesis that a minimum viable model is possible. There is also no productionized data pipeline to be delivered to the end user for consumption of actual results. 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.

Part II. Key stakeholders

1. EpowerX team
   1. Abhijeet Singh
   2. Pratik Khandelwal
   3. Rohit Singh Chauhan
2. Viteos ML team
   1. Rani
   2. Data engineering resource
   3. Machine Learning resource
3. Viteos data team
   1. Girish and team
   2. Tejas and team
4. Viteos Business team
   1. Hari
   2. Darshan
   3. Vivek

Part III. Tasks and Deliverables

Section A. 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 20) will be bucketed to test the hypothesis.
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. Iterate steps 2 – 9
   1. Explanation : Since machine learning model development is an iterative process, we will iterate steps 1 to 9 till we exhaust the bucketing techniques at hand and come up with a strategy for minimum viable model.
9. Document and deliver
   1. Explanation : For final delivery, we will deliver final code with documentation and knowledge transfer.
10. Support for existing 5 clients
    1. Explanation : Constant support + code revamp to maintain constant accuracy of 80% across clients delivered in previous phase.

Section B. Project deliverables

1. Data gathering process
   1. Viteos will give recon processes to be covered as well as list of clients EpowerX has to get the data for.
   2. EpowerX does initial data analysis to check if ML is applicable on client data. EpowerX approves clients and clients are locked for further analysis.
2. Variable analysis and selection
   1. EpowerX analyses key variables responsible for recon decision making.
   2. EpowerX shares analysis and code with Viteos ML team.
   3. Viteos ML team tests and ratifies code and analysis based on
      1. Code sanity
      2. Code scalability
      3. Version controlling
      4. Confounding variables
      5. Variable independence
      6. Variable correlation with actual output
   4. Viteos ML team shares suggestions on refactoring and changes with EpowerX team.
   5. Iterate 2a to 2d until key variables are selected for each recon process
3. Variable cleaning
   1. EpowerX cleans key variables found in 2
   2. EpowerX shares variable cleaning process and code with Viteos ML team
   3. EpowerX team writes unit tests for individual modules
   4. Viteos ML team ratifies variable cleaning code and process based on
      1. Exception management
      2. Code sanity
      3. Code scalability
      4. OOPS modularization
      5. Fixing outliers
      6. Handling missing data
      7. Handling data duplicates
      8. Data type conflicts
      9. Check if all tests pass
      10. All modules are covered by tests
   5. Viteos ML team shares suggestions on refactoring, testing and changes with EpowerX team
   6. Iterate 3a to 3d till variable cleaning is fully structured with regards to logic, exception management, code modularization and testing.
4. Client bucketing
   1. EpowerX researches client bucketing approaches applicable to recon process’ data
   2. EpowerX shares client bucketing approaches with Viteos ML team, Viteos data team and Viteos business team
   3. All Viteos teams discuss internally and share thoughts with EpowerX team
   4. Both EpowerX and Viteos come to a common ground for only the logic behind client bucketing approaches.
   5. Once agreed by both parties, EpowerX codes different approaches into actionable items for client bucketing.
   6. EpowerX shares client bucketing code and analysis with Viteos ML team.
   7. Viteos ML team ratifies client bucketing code based on following actionable items
      1. Speed of ML algorithm applied to bucket data
      2. Accuracy of ML algorithm applied to bucket data
      3. Intersection logic between different buckets
      4. Overall Logical reasoning and purpose behind bucketing approach
      5. Satisfaction of purpose by bucketing approach
      6. Scalability of code logic to incorporate additional clients into bucket. Note that this does not mean if OOPS modules are written to achieve scalability, rather if code bucketing logic is scalable. Code sanitization of bucketing approach will come later.
   8. Viteos ML team shares suggestions with EpowerX for refactoring and changes in the code base
   9. Iterate through 4e to 4h till final bucketing approach is ready for each recon process.
   10. EpowerX team and Viteos ML team share bucketing code overall results with Viteos data team and Viteos Business team.
   11. EpowerX and all Viteos teams discuss and come to agreement over client bucketing code base results
5. Client bucketing validation testing
   1. EpowerX will keep out validation data to test on which won’t be included in bucket building process outlined in 4.
   2. Based on results from 4, EpowerX will test results on validation data and present code to Viteos ML team
   3. Viteos ML team will ratify code written to test results on validation data.
   4. Viteos ML team will execute validation testing code on validation data and check for veracity of results presented by EpowerX in 5b
   5. Iterate 5a to 5d in case of inconsistencies or logical errors found in code
   6. EpowerX and all Viteos teams discuss results and signoff for Part V.5 to release payment
6. Documentation and Knowledge transfer
   1. EpowerX team to document all code
   2. Viteos ML team to check code documentation for all functions or modules. Code documentation should have inherent information on
      1. For functions
         1. Purpose of function
         2. Description of parameters used in function
         3. Data type of parameter used in function
         4. Description of return output of function
         5. Data type of return output of function
      2. For modules/class
         1. Purpose of module
         2. Inheritance of module
         3. Functions exposed in modules and functions used internally in module
   3. EpowerX team to document all analysis reports (2b, 3b, 4b, 4f) shared with all Viteos teams
   4. Viteos ML team and Viteos data team to check analysis reports.
   5. EpowerX team to document all results on actual data (1b,5b)
   6. Viteos ML team to check for sanity of result documentation on actual data.
   7. Viteos ML team would know all code shared by EpowerX team during different phases as both teams would work very closely. If they still have problems, Viteos ML team will share all areas of code they have issues in understanding the workflow of, with EpowerX team.
   8. EpowerX team will give KT sessions for 6g
7. Support + Code revamp for phase 1 delivery
   1. EpowerX team to revamp code to maintain accuracy of 80% across all clients delivered in Phase 1. Code revamp to include following areas:
      1. OOPS modularization
      2. Testing module coverage
      3. Exception management on module level instead of linear flow exception management written currently
      4. Accuracy of 80% on all existing clients.
   2. Viteos team to check and test all new code delivered for maintaining accuracy of 80% for Phase 1.
   3. EpowerX team to give KT sessions to any new hire inside Viteos for Phase 1 ML project.

Part IV. Cost, Duration and Client coverage

Section A. 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 5 X 3 = $69,600

Section B. Duration:

Total duration : 5 months

Section C. Total clients to be analysed : 20

Part V. Project Payment Cycle and Signoff criteria

Section D. Project payment cycle

1. 20% payment on project signing within first month of official start.
2. 10% on completion of 3.B.7a (Support + Code revamp for phase 1 delivery)
3. 10% on completion of 3.B.1 (Data gathering process) and 3.B.2 (Variable analysis and selection)
4. 10% on completion of 3.B.3 (Variable cleaning)
5. 20% on completion of 3.B.4 (Client bucketing)
6. 20% on completion of 3.B.5 (Client bucketing validation testing)
7. 10% on completion of 3.B.6 (Documentation and Knowledge Transfer) and 3.B.7b,c (Support + Code revamp for phase 1 delivery)