EDMCouncil

XORIANT CDi

EDMWebinar

AI/ML for Operational Data Management

Live Date: June 11, 2020

Featuring:
Masood Khatri, VP, Head of Digital Practice, Product Owner
Apnav Agarwal, Director, Data Management and Governance

Recording: bit.ly/2znNypw

Presentation: bit.ly/3hANwM7

edmcouncil.org **cdi.xoriant.com**

WEBINAR Q&A:

Where do I start when implementing this type of capability?

First, we reached out to our operations team to get a better understanding of their challenges. We learned from our conversation where the team was spending most of their time and where the data quality issues reside. We also looked into our historical data over the last few years and established a pattern with difficulties faced by our operations team. From our discussions, we selected use cases that can bring immediate benefits back to our business that can be quickly implemented. Our successful proof of concept led us to implement the use cases and brought overall efficiencies to our operations.

Could you elaborate on architecture used in these cases?

We used the following environment: Python, Tensorflow, GCP, REST API.

Do you have any use cases for producing finance or risk reports using AI/ML?

Finance and/or risk reporting are part of our roadmap.

Do you use any Data Catalog automation with ML/AI?

As of now we have not automated Data cataloging with AI/ML, but we will be taking over this use case after doing the usability and business benefits analysis.

For the address segmentation, did you use Named Entity Recognition to identify the different parts of the address?

NER was not used. We used Pre-Trained Word Embeddings, Address Element Tagging, and Recurrent Neural Networks.

Could you let us know how such implementation can be audited? What will be the functionality to enable auditing such tools?
We suggest you store 3 types of data in your systems. Original data fed to the model, Predictions returned by the model and corrections, if any, applied to the predictions data by your operations team.

Do we have any other attributes also in pipeline which will be considered to be included in the AI/ML to save time and improve the quality?
Yes, several use cases are part of our roadmap and we will work on them post completing our immediate use cases.

What is the data size when AI/ML will work well?
AI/ML can handle any data volumes. If your data is in terabytes and petabytes, you may have to utilize Big Data Technologies for AI/ML. For smaller volumes (e.g. Few tens of millions of rows), python works just fine.

How do we deal with sparse data in innovation?
It all depends upon how close your training data is to reality. The Data Science model will get trained on such sparse data and figure out how to distinguish between sparse and dense data.

How have you addressed mastering complex data sets (i.e. customer/client data that may not have authoritative data sources to validate against)?
Operations team guided by extensively researched and written data sourcing policies such as country-specific rules collects information manually and record the audit source. Validation against the recorded source and data sourcing policies help to master such data sets.

How do you make 100% automation in Data Quality?
Deep Learning models learn better as more and more training data is passed on to them. This is an iterative process to improve your accuracy

Is Xoriant also working on standardizing the data attributes like FIBO in reference data for better result in machine learning?
Yes.

Regarding the DQ use case: have you been able to check off data in terms of technical quality or even in terms of business meaning (like an DQ business rule)?
The cases where the prediction probability doesn’t give a direct answer (i.e. either Accepted or Rejected), are manually reviewed by our operations team.

What would be the typical time-period for training the models?

It depends upon the technical environment you are using, the volume of data you are feeding to the model, and the types of models you are utilizing. It could vary from few hours to few days. There are techniques available to speed up model training

How do you define accuracy when you say 90-95% accuracy is the goal?

You compare Predicted versus Actual values and measure what % of values got correctly classified.

How do you handle semi-structured / unstructured data from extraction perspective?

OCR allows you to extract printed text. There are other techniques available to extract text from Web pages. If you have to feed such data to Data Science models, Word Embeddings is one of the ways to go. There are Pre-Trained Word Embeddings available to make your life easier.

Are we handling duplicate data in this Address model?

Our emphasis was on Address Standardization till now. We plan to take up Address Correction in the next phase.

What are the key takeaways of this webinar?

We started this journey with a concept and kept working on the vision of improving data quality and gaining operational efficiency to create value for our organization and our clients. This is a result of teamwork between our operation, technology, and data science practice teams. It is not easy to get the right solution on the first try and we worked with a trying lot of solutions, models, etc. We had support from executive management that allowed us the time & resources to be successful. We would say this is a journey, so keep trying and working towards the strategic vision for your organization.