OakNorth 2020 Machine Learning Initiatives

OakNorth Machine Learning

17th March, 2020

Abstract

This document tracks down all initiatives that are driven by Machine Learning and Statistics Modelling techniques in OakNorth. Each section outlines an initiative, starting with a list of business stakeholders, brief introduction of the task, dependencies of the initiative, and potential client-facing applications that depend on it. Some will have additional section to furthermore justify the initiative. The initiatives in this document will be moved to ML JIRA board in a form of an **EPIC** and ready to be picked up by ML engineer(s) once it is approved by business stakeholders.

Each initiative is derived directly or indirectly from OakNorth 2020 vision[1]. The ordering of the sections are randomly chosen.

1 Sentiments Analysis

Keywords— Multi-class classification. Aspect based opinion mining.

1.1 Business Stakeholders

Toby Smith

1.2 Intro

Being able to derive sentiments from text with respect to a borrower, product, or an industry, allows us to understand general opinion for the entity in question, from large stream of textual data efficiently. This alternative signal is shown to be an effective indicator for market movements [2, 3], or institution performance [4, 5]. Exposing sentiments derived from news, research and product reviews is beneficial for both our customer for monitoring purposes, and credit analysts for loan assessment.

1.3 Dependencies

This initiative has dependencies on data, and optional dependencies on three other technical components.

Different types of sentiments need slightly different types of data. Specifically:

1. Subscription on Social Media feed, e.g. Twitter, Weibo (Company, Sector and Sovereignty Sentiments)

2. Subscription on News feed, e.g. Bloomberg EDF, Capital IQ Key Development. (Sector and Sovereignty Sentiments)

3. Subscription on Products reviews feed, e.g. Amazon/eBay/Yelp Products Review. (Company, Sector and Sovereignty Sentiments)

To derive company sentiments, we need **Named Entity Disambiguation**, s.t. we know what is the object for positive/negative sentiment in a paragraph. In this line of work, it is assumed that the corpus providing sentiment signals will have coverage on the companies. For news, however, it is rare SMEs are mentioned. Therefore the PI has to carefully decide what text corpus should we procure to derive SME sentiments.

To derive sector and sovereignty sentiments, we will need **Topic Classification**, so that we know what is the topic/sovereignty the text mainly concerns. Note that some vendors do provide topic classifications along with text so we may not need to build them in house.

1.4 Client Facing Applications

1. Sector Sentiments, Property Sentiments, etc.→Monitoring

2. Key Drivers Recommendation →Credit Analysis & Tooling

2 Table Understanding

Keywords— Structured prediction. Information Extraction.

2.1 Business Stakeholders

Kristjan Kaar, Daria Saulenko

2.2 Intro

As stressed in vision document, extracting tables, understanding its structures and semantics, aka, converting data items to OakNorth taxonomy, are keys to automate borrower data extraction. Whilst works from ML team on Table Extraction and Structure Understanding have shown promising results that we can automate extracting and recovering tabular structures, our production system cannot automatically understand these tables, as

1. the data items in these files are named in higher degree of variability
2. calculation relationships in the sheets are not captured.
3. there are specific data items (e.g., Revenue from Solar Power Storage in Tesla statement), and OakNorth taxonomy does not (and will not[1]) have them.

For example, in figure 1, the current system can easily ingest **Revenues**, but not items such as **Contingent legal fees** as it is a specific item that OakNorth taxonomy does not contain. Nevertheless, for verification purpose, we still want to ingest the data item. We also want to automatically ingest the fact that data item **Contingent legal fees** contributes positively to calculation of **Total portfolio operations**, and negatively to calculation of **Net revenue (loss)**. These additional requirements cannot be handled by the current system.

In this initiative, we aim at zero-basing the table understanding model we have on production, and deliver a new system that can capture unknown data item and calculation dependencies across data items.

2.3 Dependencies

This initiative has dependencies on training data, the definition on OakNorth taxonomy, and dependencies among data items in OakNorth taxonomy.

One recently available training corpus for this task is SEC iXBRL filings, which captures stylistic variations of tables in filings, representation differences on data items, and calculation dependencies among data items. An example can be seen in `https://www.sec.gov/ix?doc=/Archives/edgar/data/934549/000093454919000017/actg2018123110-k.htm`.

[1]Company specific data items are less useful as it is hard to use them for downstream tasks, such as benchmarking.

	Mar. 31, 2018
Revenues	$ 62,093
Portfolio operations:	
Inventor royalties	21,744
Contingent legal fees	15,759
Patent acquisition expenses	4,000
Litigation and licensing expenses - patents	2,989
Amortization of patents	5,330
Impairment of patent-related intangible assets	—
Other portfolio expenses	—
Total portfolio operations	49,822
Net revenue (loss)	12,271

Figure 1: An extracted table from a SEC filing. The row starts with **Contingent legal fees** are not recognised in OakNorth taxonomy, while it contributes towards the calculation of **Total portfolio operations** positively and **Net revenue (loss)**. negatively. It is beneficial to ingest those data items as well for verification purpose.

Note that, using SEC iXBRL filings as training data assumes that we can easily map OakNorth data items to US-GAAP data items.

2.4 Client Facing Applications

1. Raw Financials Ingestion →Internal Borrower Data Ingestion
2. Raw Financials Ingestion →External Data Operation & Ingestion

3 Document Classification

Keywords— Multi-class classification

3.1 Business Stakeholders

Kristjan Kaar, Jagjit Sandhu

3.2 Intro

Relation managers for each borrower are required to classify each document uploaded by borrower to corresponding bucket, so that analysts can locate necessary information during credit assessment process faster. This manual categorisation process is time-consuming and repetitive. Given the fact that we already have large amount of bucketised documents in Nebula stack, can we build a system to automatically classify this document?

3.3 Dependencies

This initiative has one data and two technical dependencies. We require the bucketised raw borrower files from Nebula, which is used to be served by an SFTP, to be available on nebula production stack. For technical dependency, we need enclave [6] to be ready, and document structure understanding to be ready.

3.4 Client Facing Applications

1. Document Upload →Credit Analysis & Tooling
2. Document Upload →Data Operation & Ingestion

4 Improvement on Peer Suggestion and Company Search

Keywords— Learning to Rank, Time Series Modelling, NLP

Value Proposition— As an RM, after borrower financials are ingested and a sector selected, I am automatically presented a list of Peers for comparison.

Success Metric— Precision@K, Mean Average Precision, Normalised Discounted Cumulative Gain.

4.1 Business Stakeholders

Supratik Shankar, Jagjit Sandhu

4.2 Intro

The recently defined successful metric for peer suggestion system require a new set of manual annotations every time when a new baseline needs to be evaluated. At the moment, the collection of annotations are done in an-hoc excel workbooks. We crossed a point that we need a more systematic pipeline to improve and automate the process.

On the other hand, with more annotations being collected, we can start investigating supervised modelling approach, e.g. Learning to Rank, to furthermore improve quality of suggested peers.

Finally, a closely related task, company search, can utilise the above annotation to improve the relevance of search results.

This initiative tracks the effort of implementing an annotation pipeline, researching on (semi-)supervised approach on improving quality of peer suggestion, ranking of company search results, and productionisation of the outcome from the research.

4.3 Client Facing Applications

1. Peer Suggestion →Credit Analysis & Tooling, Monitoring
2. Company Search →Credit Analysis & Tooling, Monitoring

5 Topic Classification

Keywords— Multi-class classification, Structured prediction

Value Proposition— For topics on borrower files: as an RM or Credit Analyst, I view an organized summary of key topics extracted from borrower documents; these are presented consistently post-ingestion within the borrower overview section.

Success Metric— Macro F-score

5.1 Business Stakeholders

Toby Smith, Morgan Williams

5.2 Intro

Classifying news, social media posts, and paragraphs in research reports or borrower submitted files to relevant topics are crucial for our clients to consume textual information in OakNorth platform. This allows them to quickly retrieve a piece of information they need, i.e. improve search engine, allow other algorithms such as sentiment analysis to correlate signals with topics such as sovereignty, or business sectors. This initiative yields

1. A topic classification library that returns a set of topic codes (with uncertainty) given parameters of domain and texts.
2. A component in the ETL pipeline that enriches texts during ingestion.
3. A new service in One-API that serves topic classification request.

The library will associate each textual unit, defined by stakeholders[2], to a set of topic codes. The set of topic codes will differ from different kind of data[3].

5.3 Dependencies

For topic classification on internal borrower files, the dependencies are

1. Internal Borrower Files
2. Definition of topics for borrower files

[2]Textual unit has to be defined according to the kind of data. For example, it does not make sense to tag a full pdf file with 89 pages submitted from borrower w/ a set of topics as each page (or even a paragraph) can have very different topics.

[3]Topic codes have to be defined differently as some topic codes do not generalise to all kind of textual data. For example, a topic code such as "Fixed Charge" can be used to associate w/ a paragraph in borrower file which describes debt structure, while it is unintuitive to be used to tag a social media post

3. Enclaves [6]
4. Document Structure Understanding

For topic classification on external data, the dependencies are

1. News/Social Media data on data platform.
2. Definition of topics. e.g. Sovereignty and business sectors that we are going to support.

5.4 Potential Topics

From decision trees created by Analysts, we found out that flagging out the following paragraphs from a borrower file will be useful:

1. Facility Amount
2. Purpose of Loan
3. Term
4. Security (Collateral)
5. Business Description
6. Geography of Business
7. Product Lists
8. Utilisation details of past loans
9. Management Information
10. Legal Information
11. Collaborator Details
12. Group Structure

5.5 Client Facing Applications

1. News Search/Alerts →Monitoring
2. Sentiments →Credit Analysis & Tooling
3. Data Operation & Ingestion

6 Content Search Engine

Keywords— Learning to Rank, Information Retrieval

6.1 Business Stakeholders

Jagjit Sandhu

6.2 Intro

Typical PDF viewer only allows users to search information by exact matching queries. As borrower submits tens of files, while each of them contains hundreds of pages, it's hard and time-consuming for an analyst to retrieve relevant information. This effort tracks the implementation of a better ranker for search engine that, based on a query in logical form (or equivalents) [7], retrieve relevant textual units in multiple borrower files.

6.3 Dependencies

This effort depends on one data dependency, and four technology dependencies.

For data dependency, we need raw borrower files available on production stacks.

The technology dependencies are:

1. Availability of Enclaves [6]
2. A search engine that indexes raw borrower files
3. Named Entity Disambiguation
4. Topic Classification

6.4 Client Facing Applications

1. Searching Raw Borrower Files →Credit Analysis & Tooling
2. Searching Raw Borrower Files →Monitoring

7 Named Entity Disambiguation

Keywords— Structured prediction, Ranking

7.1 Business Stakeholders

Morgan Williams

7.2 Intro

Named Entity Disambiguation (NED)[4] is a task of identifying what are the entities referred by sub-strings in a sentence, given a knowledge base (KB) of ground truth entities. Despite of the fact that NED is not directly client-facing, it is a crucial component for many down-stream applications, such as sentiments, content search engine, peers suggestion, etc.

7.3 Uniqueness

Having a powerful NED can improve multiple components in OakNorth system vastly. For example, credit analysts can search mentions of a specific company (or all its subsidiary companies) across all submitted files from a borrower, even if the mentions are named with aliases or, worse, incompletely. NED with KB allows us to contextualise search[8][5] such that search results can be enriched. NED can be correlated with sentiment signals so that we can infer sentiment time-series for each company.

Despite of the business values, to the authors' best knowledge, there is no commercial engine that can accurately link entities in text that mainly discusses about SME. One of the major reasons is lack of data to populate a SME centric KB to build NED, and, more importantly, NED is a notoriously difficult task [9, 10, 11, 12]. Since OakNorth platform has access on text corpus from different SME business, and we have a strong in-house ML team for building proprietary NED algorithms, we will be in a good position to overcome the difficulties and obtain competitive advantages.

7.4 Dependencies

For data dependencies, the following types of data are required on KB

1. Aliases Relationship
2. Industry Relationship
3. Company Descriptions Relationship

The following are optional data dependencies on KB:

[4] We use NED and Named Entity Linking interchangeably in this document

[5] An illustrative example is: search Obama in Google and you will see relevant entities on the right side of the windows.

1. Key Phrases/Tokens Relationship
2. Company Website URL Relationship
3. Supply Chain Relationship
4. Co-occurrence Relationship
5. Hyper Link Relationship

Apart from the data dependencies from KB, the additional data dependencies are:

1. Internal Borrower Files
2. News/Social Media data on data platform.

The technology dependencies are:

1. A knowledge graph
2. Document Structure Understanding
3. Entity Consolidation

7.5 Client Facing Applications

1. Sentiments →Monitoring
2. Content Search Engine →Credit Analysis & Tooling
3. News Search Engine →Monitoring

8 News Search & Recommendation Engine

Keywords— Information Retrieval. Learning to Rank.

8.1 Business Stakeholders

Toby Smith

8.2 Intro

While there will be millions of news flowing into the platform daily, we have to ensure that the information shown to clients is relevant, not redundant, and curtail to preferences of each individual. This initiative aims to build a search and recommendation system to allow clients to read the most relevant news with respect to their need, specified by query in logical form (or equivalents) [7].

8.3 Uniqueness

Overlaying real-time news with loans portfolio in a single dashboard is an extremely powerful feature, as this unified view helps users to quickly understand their portfolios from different sources in a single place, and a product like such with SME focus is the first of its kind in industry. Nevertheless, without a proper search engine to filter, refine, rank and recommend information, we will easily flood users with useless signals on the dashboard. This initiative is a core foundation to enable platform displaying news on dashboard. The algorithm yielded from this initiative will be an unique IP as it will be optimised specifically for SME.

8.4 Dependencies

For data dependencies, we would need News/Social media data available on data platform.

The following technology dependencies are mostly optional, in the sense that the lack of them does not block the completion of this initiative, while having them will lead to major improvements for search.

1. Topic Classification
2. Named Entity Disambiguation
3. Sentiment Analysis

8.5 Client Facing Applications

1. Monitoring

9 Key Drivers Recommendation

Keywords— Quantitative Finance. Risk Factor Analysis.

Value Proposition— The ON Platform provides explanations of returns (or other target financials) by fundamental drivers such as Macro Economic, Industry KPIs. The estimated exposures w.r.t factors can then be used to formulate warnings (so that when the company is highly exposed to KPI x, and when there are warnings to X, a warning will be triggered for the borrower)

Success Metric— Performance of Backtesting for evaluating warning triggering.

9.1 Business Stakeholders

Christopher Woon

9.2 Intro

In the process of credit analysis for SME lending, analysts are required to come up with a projection of the company performance to correctly evaluate the risk of such loan. This step, aka Financial Modelling, usually consists of three necessary stages

1. Industry Review - Analysts read through news, past research, company filings etc to understand the industry of the borrower.
2. Analysts, based on experience and the outcome from industry review, hand-pick several factors and hand-setting weights for the factors.
3. Review the fit

Key-driver selection step is time-consuming, as this is very much experience driven and the learning curve of individual analyst is steep. Furthermore, the current selection step involves with a manual data collection and mental exercise to summarise key factors from quantitative and qualitative data.

This initiative tries to build a model that can ***suggest*** the most related key drivers with respect to a target variable of a base borrower, based on historical spreads, previously selected key-drivers by senior analysts, jointly with peer financials. The target variable can be proxy metrics of company performance, or it can be covenant.

9.3 Dependencies

Here lists the data dependencies:

1. Industry Data Time Series on Data Platform

2. Company Financials

Note that, as we want to have more explaining power, the frequency of time-series should be as high as possible, e.g. daily is better than quarterly.

The following are optional technology dependencies, as they can provide potential factors to be correlated w/ target financials by the recommendation algorithm:

1. Sentiments & Named Entity Disambiguation
2. Sentiments & Topic Classification

Peer suggestion is another optional technology dependency, as it can provide a smaller estimation universe by suggesting a list of companies that is closest in terms of financial behaviours.

9.4 Client Facing Applications

1. Alerting on Abnormality on Key Drivers →Monitoring
2. Financial Modelling →Credit Analysis & Tooling
3. Suggesting news that relates to key drivers →Monitoring

10 Entity Consolidation

Keywords— Data alignment, Deduplication, Combinatorial Optimisation

Value Proposition— Clients using our platform will not see duplicated companies from different data vendors we have. They will only see one that is consolidated.

Success Metric— Precision & Recall

10.1 Business Stakeholders

Morgan Williams, Tiziano Quarantotto

10.2 Intro

As we source data from multiple vendors, there will be duplicated information flowing into OakNorth. Since OakNorth data platform is entity centric, we have to consolidate information which refers to the same entity to avoid unwanted side-effects in downstream applications. This initiative aims to construct a generalised framework that can deduplicate among entities of the same type ingested from different vendors with probabilistic model(s). To start with, we will focus on consolidating information of companies sourced from Factset, Orbis, and Capital IQ. We will then extend the framework to deal with consolidating transactions and listings of properties.

10.3 Dependencies

The data dependencies are:

1. Company data from at least two vendors.
2. Historical property listing and transaction data.

10.4 Client Facing Applications

All client facing applications, e.g. Credit Tooling, Monitoring, that retrieve entities from OneDB will be benefited vastly from this initiative.

11 Document Structure Understanding

Keywords— Structured Prediction, Deep Learning, Computer Vision, Language Modelling

Value Proposition— As a Data Analyst (at ON or a client), I have increased efficiency of ingesting financial data.

Success Metric— Time to ingest a document

11.1 Business Stakeholders

Kristjan Kaar, Ezgi Bereketli

11.2 Intro

One of the major time blockers for credit analysts have been manual entries of data from raw files submitted from borrowers. This efforts aims at semi-automate the manual ingestion process by computer vision and language modelling techniques to

1. Segment the documents into different types of contents: Tables, Paragraphs, Figures, etc.
2. Understand the tabular structure of each table, i.e. segment the table into MxN cells according to visual evidence. This include retrieval of formatting information. The system should also return how (un)certain the system is.
3. OCR (Optic Character Recognition) to obtain the contents of the tables from the pixels. OCR to obtain the texts of paragraphs.
4. Detect the topics and saliency of the tables.

11.3 Dependencies

The optional dependencies are:

1. Raw borrower files
2. Enclaves [6]

11.4 Client Facing Applications

1. Data Operation for Financials Extraction →Monitoring, Credit Analysis & Tooling
2. Content Search Engine →Monitoring, Credit Analysis & Tooling

12 Industry Classification

Keywords— Structured Prediction, Deep Learning, Natural Language Processing, Multi-class classification

Value Proposition— 1. Clients using our platform can quickly pick sectors by entering keywords. 2. Most companies in DP does not have L6 OneClass classification code. We can improve coverage by prepopulating L6 code for all companies w/ description in the platform. This improves data consistency in our platform.

Success Metric— Macro F-score

12.1 Intro

Industry classification of a company is crucial analytics for OakNorth platform user to refine company and peer search results. As some companies are not tagged with their industries, this often leads to inferior search results. This initiative tracks the effort of implementing a Machine Learning classifier that infers industries code given company descriptions. Specifically, it yields

1. A library that infer industries (with uncertainty) given company descriptions.
2. A new component in ETL pipeline that automatically tags companies with their industry code.
3. A new service in One-API that serves industry classification requests.

12.2 Dependencies

This effort depends on the availiability of company dataset from Factset.

12.3 Client Facing Applications

1. Automatically suggest industry code in credit bench when borrower is filling in a new application.
2. Peer Suggestion →Credit Analysis & Tooling, Monitoring
3. Company Search →Credit Analysis & Tooling, Monitoring

13 Information Extraction from Property Listings

Keywords— Computer Vision, Deep Learning

13.1 Intro

To evaluate a property loan, credit analysts needs to conduct a thorough analysis on the *similar* properties of the area to verify the hypothesis suggested by the borrower. To measure similarity across different properties we have to extract information from unstructured listing data to aid the analysis, as they are not provided in standard feeds of property data vendors, while the premium feed is extremely expensive [6]. The initiative tracks the effort of extracting crucial attributes of a property from unstructured data such as the webpage, floor plans or other attachments of the listing. Here lists the set of attributes analysts are interested in:

1. Square Footage
2. Surrounding amenities
3. No. of bedrooms, living rooms, gardens, parking, garage, lift, gated community, concierge services, retirement property.
4. whether the property is New Build, Old Build, Refurbished)
5. Property Type (Detached, Semi-detached, Terrace, End of Terrace, Bungalow)
6. Sub-classification (1st/2nd Floor, lower ground floor, mezzanine floor, penthouse)

13.2 Dependencies

This initiative depends on the availability of historical property data. There is no known technology dependency.

13.3 Client Facing Applications

1. Property Heat Map →Monitoring, Credit Analysis & Tooling

[6] Reach out to business stakeholders for the estimation of cost.

References

[1] OakNorth. Oaknorth platform 2020. 2019.

[2] Bloomberg. Trending on twitter social sentiment analytics.

[3] Xin Cui, Daniel Lam, and Arun Verma. Event-driven feeds quantitative research - embedded value in bloomberg news and social sentiment data.

[4] Walter Kasper and Mihaela Vela. Sentiment analysis for hotel reviews. In *Computational linguistics-applications conference*, volume 231527, pages 45–52, 2011.

[5] Tim Loughran and Bill McDonald. Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54(4):1187–1230, 2016.

[6] Iat Chong Chan, Dimitar Popov, Konstantinos Perifanos, Sandeep Agarwal, and Sean Hunter. Privacy-preserving data platform.

[7] Robert May et al. *Logical form: Its structure and derivation*, volume 12. MIT press, 1985.

[8] Nikos Voskarides, Edgar Meij, Ridho Reinanda, Abhinav Khaitan, Miles Osborne, Giorgio Stefanoni, Prabhanjan Kambadur, and Maarten de Rijke. Weakly-supervised contextualization of knowledge graph facts. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 765–774. ACM, 2018.

[9] Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. Robust disambiguation of named entities in text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 782–792, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics.

[10] Heng Ji, Ralph Grishman, Hoa Trang Dang, Kira Griffitt, and Joe Ellis. Overview of the tac 2010 knowledge base population track. In *Third Text Analysis Conference (TAC 2010)*, volume 3, pages 3–3, 2010.

[11] Zhaochen Guo and Denilson Barbosa. Entity linking with a unified semantic representation. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 1305–1310. ACM, 2014.

[12] Amir Globerson, Nevena Lazic, Soumen Chakrabarti, Amarnag Subramanya, Michael Ringgaard, and Fernando Pereira. Collective entity resolution with multi-focal attention. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 621–631, Berlin, Germany, August 2016. Association for Computational Linguistics.