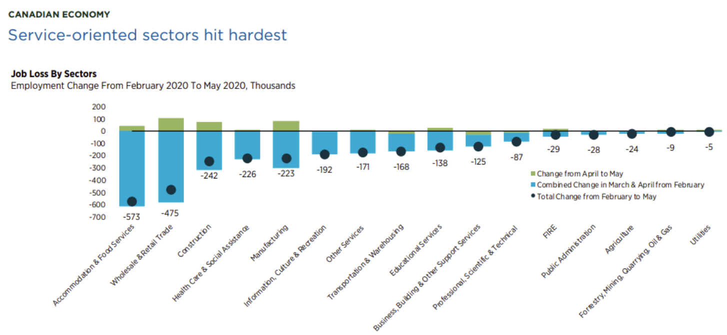
Capstone Project Proposal

Chromilo Amin | Wednesday, July 16, 2021 | Bertelsmann Technology Scholarship program

Business Goals

Project Overview and Goal What is the industry problem you are trying to solve? Why use ML/AI in solving this task? Be as specific as you can when describing how ML/AI can provide value. For example, if you're labeling images, how will this help the business?	Given the COVID pandemic is unlike any other pandemic, service-oriented jobs are the most heavily affected in the coming months and we will be using AI/ML to predict the effect. We will use the following BentallGreenOak company prospectus as a reference guide to start with CAN EN: https://bit.ly/2Y8Bulz bgo-perspective-canada-mid-year-2020 (Optional) US: https://bit.ly/3hDB8v3 This will solve the problem decision-makers have in deciding which real-estate sectors will be affected with a pandemic of this magnitude. It will help BentallGreenOak with their strategic 3-5-year roadmaps and be able to budget on tactical projects in the short-term (1-2 years).
Business Case Why is this an important problem to solve? Make a case for building this product in terms of its impact on recurring revenue, market share, customer happiness and/or other drivers of business success.	The service-oriented jobs highlighted in the graph below show an employment change from February to May 2020. It is important to predict how the numbers change in the coming months because of its direct impact to real estate business.

<table>
<tr>
<td></td>
<td>



CANADIAN ECONOMY

Service-oriented sectors hit hardest

Job Loss By Sectors

Employment Change From February 2020 To May 2020, Thousands

By building this AI/ML product, we will be able to reliably predict the relationships between the different sectors and provide R&D teams with ammunition to develop so recommendations to decision-makers at BentallGreenOak that could allow them to continue collecting recurring fees.

The graph above for example shows that as demand in wholesale and retail trade goes down, the food and accommodation services also decrease. This is seen by reduction in employment at those sectors. This could suggest a move to online transactions and food delivery services. BentallGreenOak could provide a limited break in monthly rent but supplement additional online services and delivery innovations for subscription fees to provide continued revenue for both parties. It may not correct the change in employment but new hires to support these new innovations will balance it out. Result is happy customers ultimately.

</td>
</tr>
<tr>
<td>

Application of ML/AI

What precise task will you use ML/AI to accomplish? What business outcome or objective will you achieve?

</td>
<td>

ML/AI will be used against job loss data from previous pandemic and epidemics and generate a model to predict sectors that will likely be affected. The data will be gathered from:

- AIDS pandemic and epidemic: 1981-present day
- H1N1 Swine Flu pandemic: 2009-2010
- West African Ebola epidemic: 2014-2016
- Zika Virus epidemic: 2015-present day

The ML/AI model will allow you to automatically:

- Fix the column feature for “NAICS”, “Geography”, and “Job Vacancy Statistics” correctly by tokenizing words to make sure the data collected since 1981 to today are all categorized similarly.
- Fill in missing values for column feature North American Industry Classification System (NAICS) codes.
- Fill in missing values for data quality column feature using the 11 different possible values and exclude rows which are below threshold.
- Exclude feature columns with no reported

</td>
</tr>
</table>

	employment statistics. • Predict a new label column called “Survive” with a percentage accuracy for sectors that will survive or not given above input. The business outcome or objective is a better understanding of which job sectors and therefore which real estate market to focus our energy on.

Success Metrics

Success Metrics What business metrics will you apply to determine the success of your product? Good metrics are clearly defined and easily measurable. Specify how you will establish a baseline value to provide a point of comparison.	The company prospectus is generated twice a year for the Canada and US. The report will provide a definitive roadmap for the Canadian prospectus, with statistics to back it up for the public and fund investors as the target audience. Success metrics we looked at will be financial KPIs (revenue) and customer satisfaction KPIs (employee engagement survey). The financial KPI can be baselined initially and if there are no negative changes in AUM (assets under management), this product has done its job. The employee engagement surveys are run annually to measure employee satisfaction at their jobs at BentallGreenOak. The baseline can be measured at the start and if the percentages remain positive, this is considered a success for this metric. This data should allow consumers to clearly make intelligent decisions on their investments and steer the company to a correct trajectory in terms of which Canadian economy and which real estate market to focus their attention on.

Data

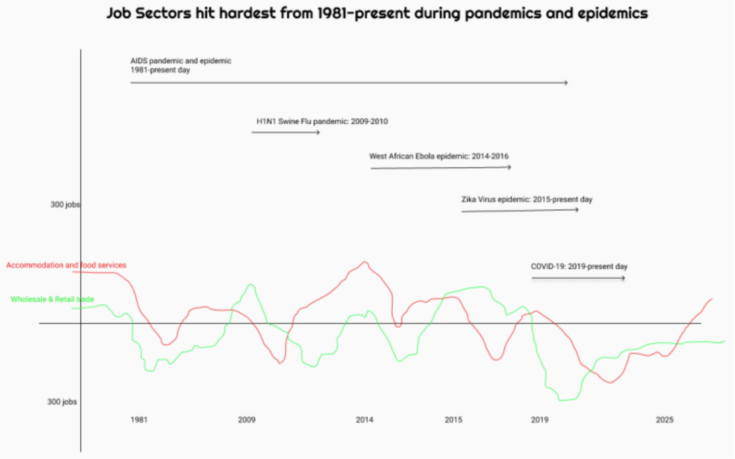
Data Acquisition Where will you source your data from? What is the cost to acquire these data? Are there any personally identifying information (PII) or data sensitivity issues you will need to overcome? Will data become available on an ongoing basis, or will you acquire a large batch of data that will need to be refreshed?	Statistic Canada open data website will allow download of data from the past 11-years which will allow us to compare job losses during the West African Ebola epidemic (2014-2016) and Zika Virus epidemic (2015-present day). To get older data for the AIDS and H1N1 pandemic, the data will have to be requested. The dataset is “Number of job vacancies, labour demand and job vacancy rate by North American Industry Classification System (NAICS), last 5 years”. https://open.canada.ca/data/en/dataset?q=jobs&organization=statcan There are no PII or data sensitivity issues as these are all publicly available. The data is static and will not require any constant refreshes.
Data Source Consider the size and source of your data; what biases are built into the data and how might the data be improved?	The zip files are about 2MB in size and requires no website registration or enrollment to download. 14100225-eng.zip There is an exclusion bias built into the data source. The dataset only includes job vacancy statistics from industries classified under North American Industry Classification System (NAICS). Other service-oriented employment like IT or Data Analytics for example are unclassified. Those other excluded classification codes make up BentallGreenOak’s real-estate business and could play a part in accurately predicting our chosen labels. To better understand if these are important, bring in subject matter experts who can better indicate which features are important to aggregate or include in the model. There is also a measurement bias where the text may vary between the different data sources collected over the years, e.g. the AIDS pandemic data as old as 1981 will have different values used for “NAICS”, “Job Vacancy Statistics”, and “Geography”. To fix this, AI/ML can be used to tokenize the words used in these fields and allow them to be automatically normalized. This will improve quality control of data collected.
Choice of Data Labels What labels did you decide to add to your data? And why did you decide on these labels	The following metadata below are already included in the downloaded dataset: • Geography • Job Vacancy Statistics

versus any other option?	- NAICS - Status - Value In addition to the above, we will initially include a “Survived” label as a percentage predicting survivability of that NAICS sector during that pandemic. A weakness for this label is that the average across all the pandemics may not make sense. There are feature columns missing that could distinguish one type of pandemic over another, providing additional data quality to this “Survived” label. Adding another label as a weighting for “Survived” label might make sense but we can add that in when we see the initial trained set.

Model

Model Building How will you resource building the model that you need? Will you outsource model training and/or hosting to an external platform, or will you build the model using an in-house team, and why?	Open-source software will be used to build the model. We will use the sequential model with Keras deep learning API written in Python and running on ML platform Tensorflow. The training of the model will be hosted on the cloud using Google Colab notebooks and will be built by our in-house R&D team because - there is no budget for the project. - open-source software using TF is easy to develop. - the business outcome is a company prospectus PDF which is only published twice a year. I would remove a job sector from the dataset but keep the other job sectors when compiling the model to train it and the impact the pandemic/epidemic has in relation to the other job sector labels.
Evaluating Results Which model performance metrics are appropriate to measure the success of your model? What level of performance is required?	The confusion matrix and F1 score will be used to measure the success of our model. A 0.5 performance score is sufficient to produce a roadmap.

Minimum Viable Product (MVP)

Design What does your minimum viable product look like? Include sketches of your product.	Figma.com has a free-tier subscription that can be used to generate a wireframe or design of the graph showing job loss/rebound predictions into 2025. 
Use Cases What persona are you designing for? Can you describe the major epic-level use cases your product addresses? How will users access this product?	We are designing based on these three types of personas: • Goal-directed personas. These are typically top-level executives and board of directors who will use this product to set company-wide goals and mission statements at BentallGreenOak. They can be Steering Committees responsible for approving budgets. • Role-based personas. These could be external auditors like Deloitte or KPMG who are specifically subcontracted to provide guidance on roadmaps. They are the R&D (Internal Research and Development) team who provide solutions to leadership when markets shift. • Engaging Personas. They can be public investors, shareholders, mutual fund RRSP investors, and pension plan holders who watch their investments and make allocations based on these product findings. They can be employees who can better align their skills to sectors that have a higher chance of survival during a pandemic. These are major epic-level use cases: • As a top-level executive or a member of the board of directors for BentallGreenOak, this product report that shows the sectors most likely to be affected by the COVID pandemic will be used to restructure and reorganize the staffing for those sectors to continue to receive recurring

	rent and revenue. It will allow them to prioritize funds to surviving sectors and innovate to help sectors heavily affected. - As an employee with a registered RRSP plan, this product report that shows the sectors most likely to be affected by the COVID pandemic will be used to change fund allocation to stronger markets that will survive the pandemic as predicted by the model. For example, funds that contain residential real estate and property management will provide higher yield as workers isolate and work remotely from their homes. Funds for residential building developments will liked do poorly so reduce those allocations. The company prospectus is readily available online on the company website as a PDF.
Roll-out How will this be adopted? What does the go-to-market plan look like?	Introduce the business proposal in the project pipeline for initial discovery intake. Once approved, meet with sponsors from the R&D department to get effort estimates and any funding needs. The development and implementation will be addressed in sprints to ensure value is quickly realized and that the finished outcome is approved for disbursement to the general public every 6 months.

Post-MVP-Deployment

Designing for Longevity How might you improve your product in the long-term? How might real-world data be different from the training data? How will your product learn from new data? How might you employ A/B testing to improve your product?	The dataset is downloaded from Statistics Canada every 6 months and the model is re-trained each time with additional rows added. This can be done indefinitely until following the same process. To improve on the real-world data additional features could be added like 1) a COVID variant; 2) drop in public confidence in the vaccine; 3) multiple waves that followed; 4) government response via benefits/stimulus. To employ A/B testing, I would establish a ground truth (or controlled test) for the existing dataset first. I would find additional datasets for the same time periods from other sources outside of Statistics Canada including 1) when the respective vaccines where release, 2) if there

	were additional waves that necessitate multiple government packages, or, 3) any corrosion of public confidence that may have slowed full public inoculation. After the ground truth is established and the new model is trained, I would use 20% of the training dataset as test calculate the performance of the model and reiterate epochs, activation functions, loss functions, and other metrics to improve on the accuracy.
Monitor Bias How do you plan to monitor or mitigate unwanted bias in your model?	To ensure there is no bias, the additional dataset will include many features or labels that may or may not contribute to changes in those job sectors. The sources may have to come from a reputable feed.