MIE 1624 FINAL PROJECT

March 31, 2020

Prof. Oleksandr Romanko
Group 15

Mostafa Kouchakzadeh, Zhaohui Qu, Zichuan Wang, Sam Weinberg

Introduction

This document contains three separate parts: MIE 1624 Curriculum Redesign, UofT Master's in Data Science and AI Program Design, and EdTech Startup Proposal. Decision are made based off analysis from a variety of sources including Kaggle surveys and web-scraped job postings. All required visualizations are contained within each section to communicate critical findings from the results and summaries of the core curriculum designs.

MIE 1624 Course Curriculum Design

In the first part of our project, we redesign the curriculum for MIE 1624, Introduction to Data Science and Analytics, based on the skills that are in demand and correlate with high salaries. We utilized Kaggle data scientist surveys from 2018 and 2019 as well as web-scraping job opportunities to extract data for analysis.

We start with the Kaggle surveys, where we use encoded salary buckets as the target variable for our analysis. We investigated which skills have a more positive correlation with higher annual compensation. Currently in MIE 1624, Introduction to Data science and analytics, we are focusing on Python as the programming language, however, we need to explore the results of the surveys thoroughly to be sure of our choice. First, we cleaned our data since both datasets have a significant amount of noise associated with them. We focused only on Canada and the United States for this report, since the course is being taught in Toronto. Figure 1 exemplifies how incorporating data from other countries can affect the analysis results. As we can see, the U.S.A. has almost a normal distribution of salary buckets whereas India's salary distribution is skewed toward lower salary buckets.

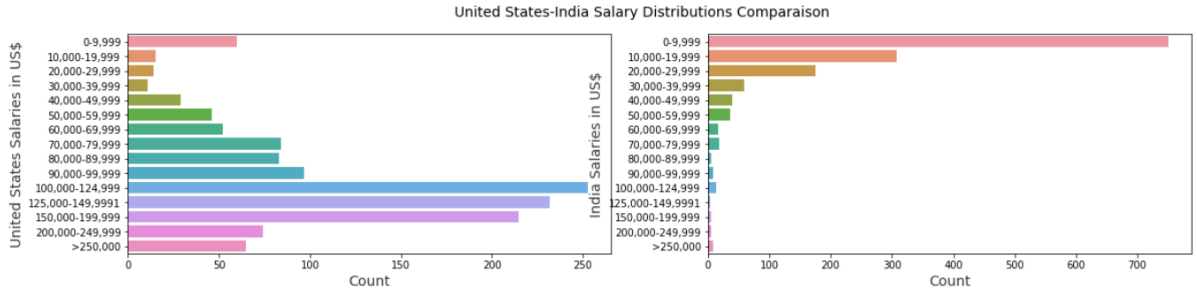

FIGURE 1- USA VS INDIA SALARY BUCKET DISTRIBUTION

Many factors contribute to a higher salary other than skills, such as years of experience, age, gender, etc. However, in the scope of this project, we are focused on skills to teach as opposed to these other factors. To start, we explored the 2018 dataset and compared the results to the 2019 results from assignment 1. We were interested to observe the differences in high-demand skills from one year to the next. We started by analyzing the programming languages that were most common and highly correlated with salary. As mentioned before, currently in MIE1624 the focus is on Python as the primary programming language. The results of the 2018 Kaggle survey in Figure 2 indicate that although Python is used for almost all the salary buckets, SQL and R seem to be useful tools to achieve higher annual compensation as well.

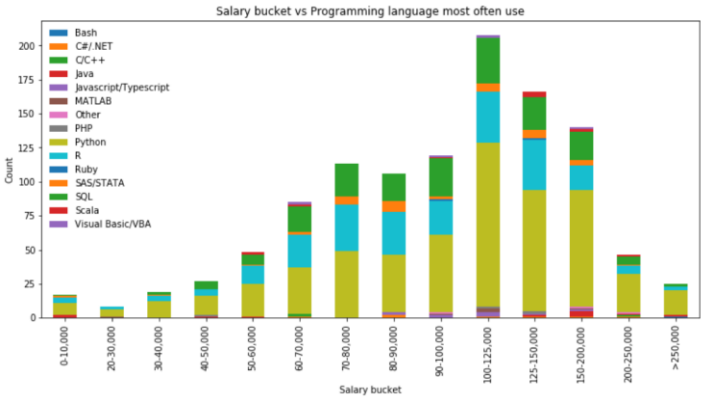

FIGURE 2- PROGRAMMING LANGUAGE BASED ON SALARY BUCKETS 2018

In the case of the 2019 survey, there was not a specific question that asked about the primary programming language which a data scientist used. We decided to use LASSO regularization to see which programming languages had a more positive weight, indicating the importance of that language. Interestingly, we found similar results in the case of the 2019 survey. Results demonstrated that SQL and R are useful programming languages if a data scientist plans to achieve a higher salary.

Next, we considered the question which asked the primary tool that data scientists used and how it correlated with higher salary. This question was repeated for both years. In the 2018 Kaggle survey, we infered that Cloud-based data software and APIs are mainly shown in higher salary buckets. In addition, Local or hosted development environments such as Rstudio or Jupyter notebook are presented in almost all the salary buckets and have the highest frequency. Similar results were found in 2019.

From other components of the analysis, we found the following skills to correlate highly with salary: Building prototypes to explore applying machine learning to new areas, IBM Cloud Analytics Engine, big data analytics products and working with tabular data. These will all be incorporated into the course with emphasized importance.

In conclusion, salary-wise analysis of Kaggle surveys from 2018 and 2019 indicated that in addition to Python, a data scientist should be familiar with SQL and R. Moreover, the redesigned course will focus more on cloud-based software and APIs, since they contribute to a higher salary than other tools. The size of each circle in Figure 3 shows the number of occurrences of each specific programming language in the average salary. We can see that again Python, R and SQL are the most prominent programming languages in the Kaggle surveys.

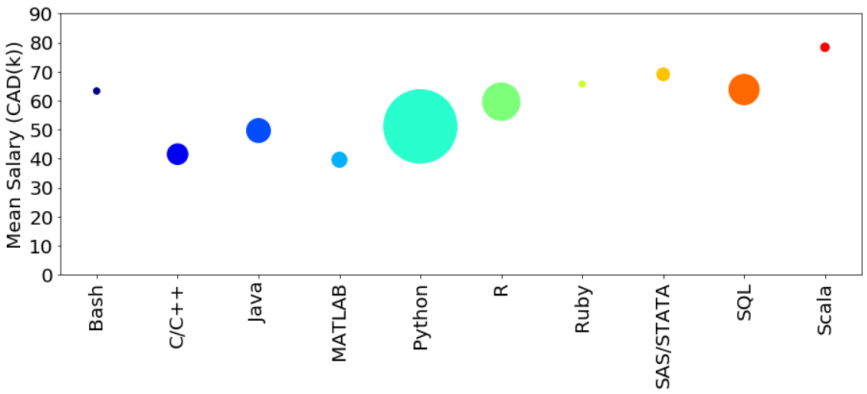

FIGURE 3- PROGRAMMING LANGUAGE BASED ON AVERAGE SALARY

In the second part of our analysis, we worked on job-posting data that was extracted from Glassdoor. By web scraping, we were able to obtain roughly 3000 job postings in the field of data science and analytics from the Glassdoor website. Unlike the last section where we analyzed skills based on correlation with salary, in this section we evaluated skills from the perspective of job opportunities. After this course, the goal is that students should be able to at least apply for internship positions in the data science field. Thus, these postings explain to us what main skills data science firms are looking for from their applicants.

Similar to what we had in case of Kaggle's surveys, most of the job postings required SQL as the programming language. In the second position, we have about 1000 job postings that require Python, and at the third place we have about 750 job postings that published R as their required programming language. However, many of these postings state multiple programming languages. In the case of MIE 1624, since the course name is stated as "**Introduction** to Data Science and Analytics", it is better to use Python as primary programming language. This choice is made due to the fact that many of the students are more familiar with this programming language because it is one of the most user-friendly environments for coding. However, we suggest that as supplementary work, the class would have several tutorials regarding SQL and an assignment that must be completed with SQL as the programming language. This would engage and familiarize students with the programming language that seems to have a high demand in the field of data science and analytics. as demonstrated in Figure 4.

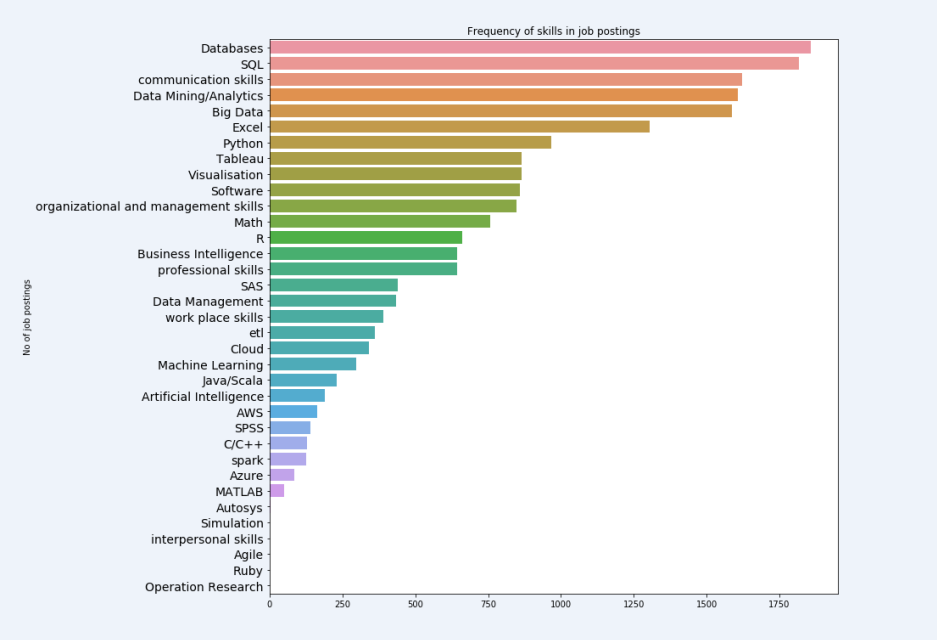

FIGURE 4- REQUIRED SKILL FREQUENCY BASED ON GLASSDOOR JOB POSTINGS

As discussed before, Python and SQL are the most important programming languages in data science. In addition, from Figure X we can interpret that working with databases in general is one of the most popular skills that job postings are seeking for. Thus, from the analysis, we are interested in adding concepts of database management and querying into the curriculum. Communication is a critical skill that a data scientist should have to be able to express analysis results to a general audience. This will be incorporated into the course through group work and presentations. Data mining/analytics seems to be quite important in the field of data science and is already a core component of the course that will remain unchanged. In this course, we would also like to add an introduction to big data tools which is a highly demanded skill. Tableau, which is visualization tool, is a required skill in many job postings. Cloud and machine learning are among the skills that were asked in more than 250 job postings.

In addition, we performed analysis using another dataset that we were able to find on the Kaggle website. It includes more than 10000 job postings including internship, full time, contractor and part time jobs. For our

analysis, we focused only on full time jobs. From the results that we had from this data frame and the previous analyses, we decided how to change the structure of the course.

Machine learning, analytics, statistical analysis, artificial intelligence are the most important features that are represented in the job postings as displayed in Figure 5.

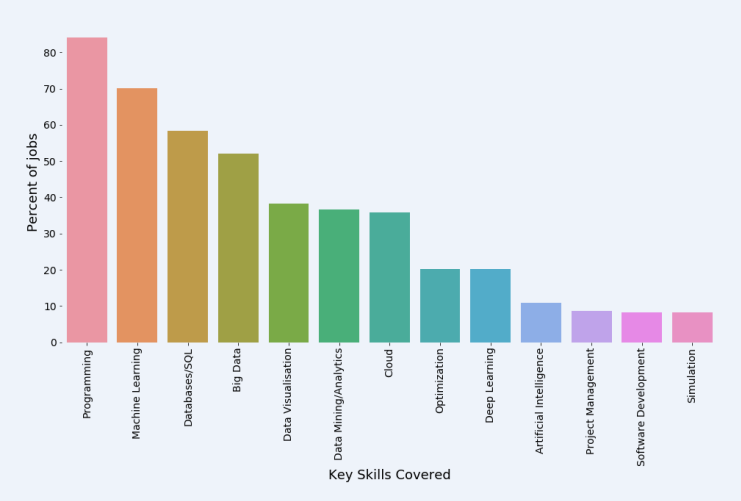

FIGURE 5- KAGGLE'S JOB PSOTING GENERAL REQUIRED SKILLS

Based on the analysis on the Kaggle's job posting data frame, programming and machine learning algorithms should be taught in the course. Database and SQL is not emphasized in the current course and will have at least two lectures devoted to it in the redesigned curriculum. We would also like to add a lecture with an introduction to big data. Data visualization should be implemented in the course project and assignments as we have in the current curriculum. Cloud computing also is a topic that can be covered in the form of an assignment or small project to familiarize students with common tools such as Azure. We have recommended that deep learning be removed from this course since it requires several lectures to obtain a meaningful introduction and is not as demanded as database skills or cloud services. The following visualization takes the results from all our analyses and combines them into one coherent redesigned course structure.

FIGURE 6: MIE 1624 PROPOSED COURSE CURRICULUM

Master of Data Science and AI Program Curriculum

Data Science and Artificial Intelligence are undergoing a phenomenal evolution, transforming the business sectors as well as technical sector. As organizations adopt and invest in data science and artificial intelligence technologies, a new style for business and management is needed – one that that pairs a leader's vision with a scientist's mastery over a growing body of specialized knowledge.

The second part of this project outlines the design of a new Master of Data Science and AI (M.D.S.A) program at U of T. The structure of the program is heavily based off the findings from the first part, as well as comparable programs from reputable universities. These include Rotman's Master of Management Analytics and Queens' Master of Management in Artificial Intelligence, among others. The curriculum is broken down into two streams: business-oriented and technical. Both streams are required to take four core courses that are essential for anyone pursuing a career in data analytics and AI. The streams then split off into advanced courses specific to the respective stream. The program culminates with a final project that matches one technical stream student, one business stream student, and a partner company to solve a real-world industry problem.

The goal of this Masters program is to provide data scientists with the skills that are currently in demand within industry. From part 1, we were able to understand which technical skills, soft skills, programming languages, and AI techniques were correlated with data scientists in high salary positions. To design this program, we also need to understand how courses should be organized to combine the learning of similar skills that synergize. This will allow students to understand how analytic techniques are used constructively to solve real-world problems. We web scraped over 5000 full-time jobs and extracted the skills that were required for each job. We then performed hierarchical clustering to visualize which skills were asked for in conjunction with one another. This analysis gives insight to how course structures should be taught to optimize the learning of synergetic skills. The following diagram demonstrates the two subsections of the hierarchical clustering analysis.

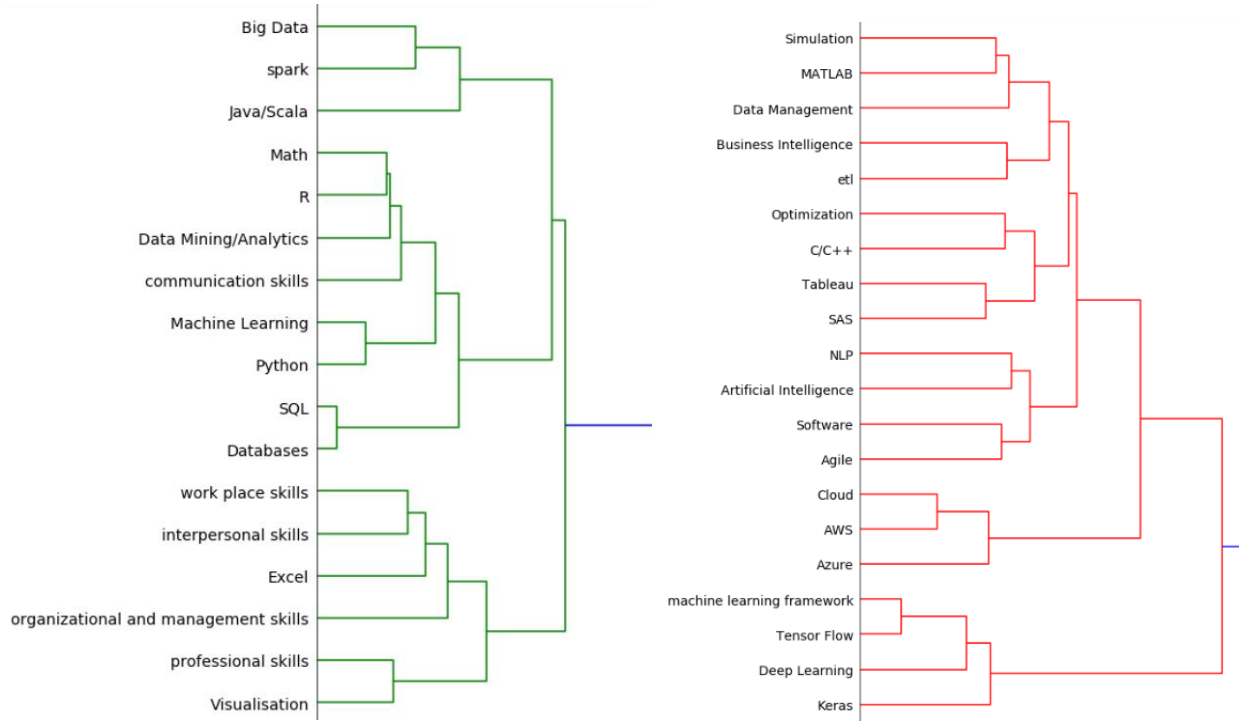

FIGURE 7: HIERARCHICAL CLUSTERING ANALYSIS OF FULL-TIME JOBS

Machine learning and AI rely on strong mathematical backgrounds in linear algebra, statistics, and calculus. Similar to Schulich and Queens, this program requires students to complete multiple Udemy bootcamps prior to the first semester to adequately prepare themselves for the subsequent courses. These bootcamps include Udemy Calculus, Statistics, and Computer Science.

The core courses are strongly based off the hierarchical clustering analysis. The core courses that are extracted from the clustering include: Introduction to Data Mining and Analytics (Python Machine Learning, Analytics), Database Fundamentals (SQL, Databases), Big Data and Cloud Computing (Big Data, Spark, Scala, AWS, Azure), Data Management and Visualization (Tableau, BI), and Neural Networks and Deep Learning (TensorFlow, Keras). These core courses represent skills that are often requested together in job postings and are in high demand. They are courses that are mandatory for both the technical stream and business stream.

Core Courses	Description
Introduction to Machine Learning and Analytics	This course is a general introduction to AI and machine learning algorithms. The course includes a breif review of statistics and calculus required for data analytics. Students will learn data cleaning, exploration, and visualization using fundamental Python packages such as Pandas, Seaborn, and Matplotlib. Prelimnary machine algorithms are covered including supervised learning algortihms (linear regression, logistic regression, decision trees, and random forests) and unsupervised algorithms (K-Means Clustering, Hierarchical Clustering, etc.).
Database Fundamentals	Databases are critical to all organizations for storing and managing large quantities of data related to different aspects of the business. Students will learn how to formulate complex queries in SQL to extract meaningful data for analysis. In addition, the course will cover web-scraping, HTML, APIs among other data mining techniques.
Data Management and Visualization	Communicating analysis results is crucial to relay information through a company with differing levels of technical knowledge. Data visualizations deliver powerful and concise summarys of technical analysis to a general audience. In this course, students will learn how to visualize data using Python and R. In addition, topics will cover dashboard technologies such as Tableau and Power BI, and presentation/communication skills for data analytics.
Big Data and Cloud Computing	In recent years, the amount of available data has exceeded are ability to process and analyze it. Distributed computing technologies enable companies to manage increasing volumes of data. In this course, students will learn the fundamentals of distributed computing using Hadoop, Spark, and Hive. The course also covers popular cloud computing services such as AWS, Azure, Google Cloud, and Watson IBM.

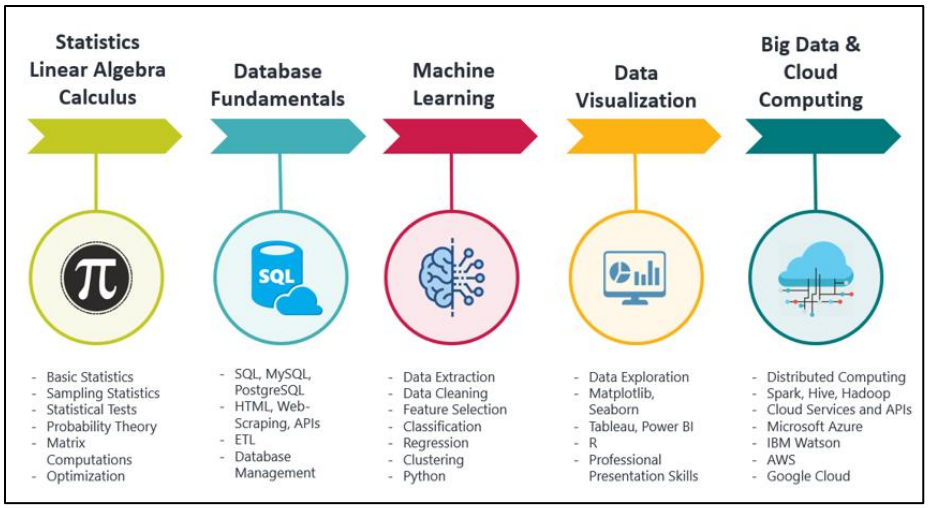

FIGURE 8: CORE COURSES FOR MASTERS PROGRAM

The technical stream courses include skills that are more advanced and build off the fundamentals learned in the core courses. They are inspired by both the hierarchical clustering and from interesting course topics that are prominent in other top programs. These courses include Natural Language Processing, Reinforcement Learning, Simulation and Optimization Modelling, Neural Networks and Deep Learning, and Times Series Analysis. Students are required to take 4 out of the 5 courses available.

Technical Courses	Description
Natural Language Processing	This course focuses on Natural Language Processing (NLP) techniques to extract and analyze textual data from a variety of sources. Students will perform NLP using Python's nltk package and various cloud services to derive real-world business insights with unstructured textual data. Topics include using APIs to collect useful data, sentiment analysis, TF-IDF, N-Grams, and other NLP techniques.
Reinforcement Learning	Reinforcement learning is at the forefront of machine learning. It makes use of rewards/penalties to incentivize actions that achieve an optimal outcome. This course will teach the basics of Markov Decision Processes (MDP) and fundamental algorithms such as Q-Learning among others. The focus will be on applying these techniques to solve real-world industry problems.
Simulation and Optimization Modeling	Simulation and optimization are powerful tools that enable businesses to make data empowered decisions on a variety of topics. Students will learn Monte-Carlo methods, stochastic and convex optimization, and graph theory. They will then apply these techniques to real-world business problem such as sales forcasting, advertisement campaigning, and supply chain logistics, and derive valuable insights to make business decisions.
Neural Networks and Deep Learning	Neural networks and deep learning have made significant breakthroughs over the past decade with the increase in computational ability. This course covers the theoretical and practical fundamentals of neural networks using TensorFlow, Pytorch, and Keras. Students will learn about the different variants of neural networks such as Recurrent Neural Networks, Convolutional Networks, Long-Short Term Memory Networks and many more. Emphasis is placed on applying these techniques to solve current industry challenges.
Time Series Analysis	Time series data is present in many real-world sectors from finance to retail. The ordered sequence of data requires special treatment when performing analysis and is still a prominent area of research in data science. Students will learn classical time series analysis models such as ARIMA, and also how to apply machine learning algorithms to time series data.

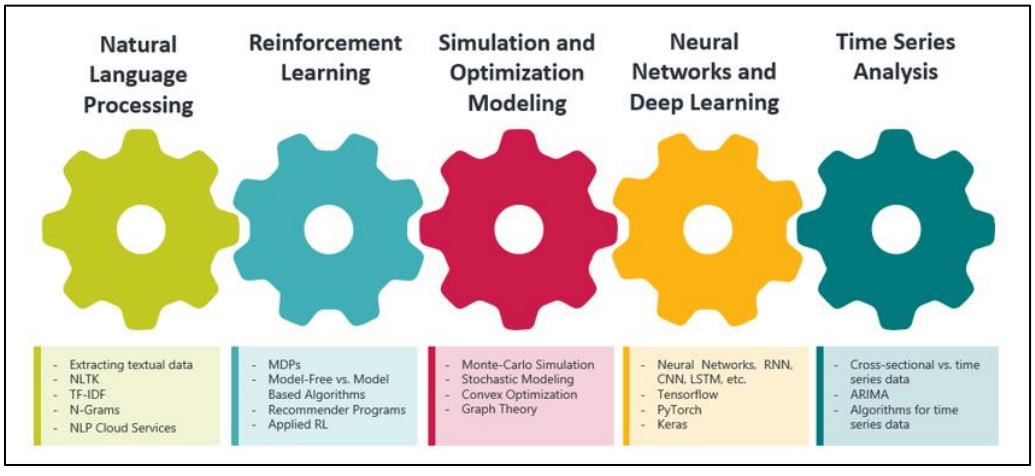

FIGURE 9: TECHNICAL STREAM COURSES FOR MASTERS PROGRAM

A successful business builds on five essential business pillars including marketing strategy, finance & accounting, supply chain & logistic, risk management and ecommerce. These five business pillars align with the strongest marketplace demand for business and management analytics professionals.

The immersive curriculum of our MBMA program offers a solid technical foundation of data analytics and machine learning and provides a guide on how to use them to achieve important business objectives. There are 5 courses for the business stream including Analytics for Marketing Strategy, Analytics for Finance and Accounting, Analytics for Supply Chain and Logistics, Analytics for Risk Management, Analytics for Ecommerce at Scale. Students are again required to take 4 out of the 5 courses.

Technical Courses	Description
Analytics for Marketing Strategy	This course will teach students how to use data analytics and machine learning techniques, along with real-world data, to achieve important marketing objectives such as: effective market segments construction; effective product design and positioning; effectiveness advertising; effective evaluation of price elasticity and brand value.
Analytics for Finance and Accounting	This course will teach students how to take advantage of the rich accounting and finance dataset to help businesses solve various problems including predictive sales analysis, client/product profitability analysis, cash flow analysis, shareholder value analysis.
Analytics for Supply Chain and Logistics	Analytics for Supply Chain and Logistics will focus on identifying, creating, and implementing effective analytics models and tools to solve typical supply chain & logistics management problems, including service process design, network design, assortment and price optimization, inventory management.
Analytics for Risk Management	This course will teach students how to combine advanced tools and techniques with expertise in strategy formulation and organizational transformation to help clients optimize their risk exposures through the use of operational risk & fraud analytics, credit risk modeling, institutional investment analytics, insurance analytics.
Analytics for eCommerce at Scale	This course will teach students how to create a KPI hierarchy to prioritize your efforts, and how to optimize as well as visualize the metrics that matter most. Some important metrics to track and optimize over are customer life-time value, customer acquisition cost.

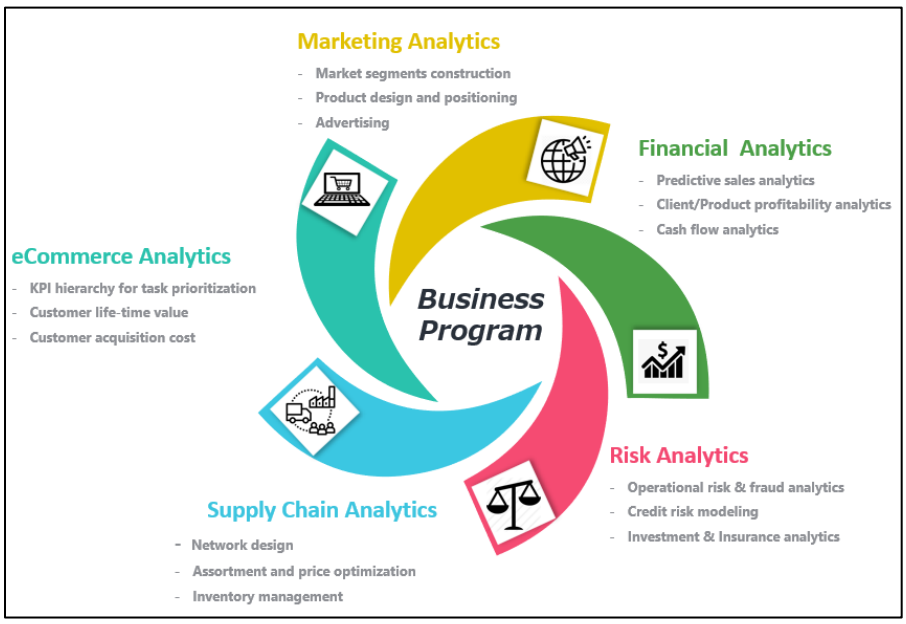

FIGURE 10: BUSINESS STREAM COURSES FOR MASTERS PROGRAM

EdTech Startup Proposal

Insight Data Science aims at answering the fast-growing demand of the job market for new skills and competencies related to advanced fields of Data Science and Artificial Intelligence and relevant for Information Technology, Finance, Insurance as well as Academic Research. It is a program that matches one business student, one technical student and a partner company to collaborate on an open-ended business problem using the power artificial intelligence.

The startup is inspired by the proposed Masters program culminating project outlined in the previous part, but the goal of the company extends beyond just the U of T program. Insight Data Science aims to provide elite graduate and undergraduate students from universities around the world with the opportunity to work on an industry project. The projects are an excellent way for students to gain valuable and practical real-world experience to bolster their resume.

The program works as follows:

- Students fill out a profile on our app or website outlining their areas of expertise and industry interests
- Our program collaborates directly with universities to obtain the students marks and references from their Professors or colleagues
- Companies provide a detailed summary of a current industry problem that they would like to solve or explore the potential to solve with artificial intelligence
- Based on the requirements of the problem, Insight Data Science will match students with industry problems that coincide with their skills and interests
- Companies can then host interviews with the matched students to select the top candidates
- The goal is to have at least one project for every student

Insight Data Science has enormous benefits for all the stakeholders. For the student, instead of spending numerous hours applying for internships or research projects, the student fills out a profile and Insight's smart-matching technology does the rest. When accepted to a project, the student gets a valuable opportunity work in a team environment on a real-world industry problem. The experience gained will aid the student in obtaining a full-time job upon graduating, as well as offer an excellent opportunity to make an impression on a prospective employer.

For the company, Insight Data Science allows them to propose a project that they are interested in solving using AI, but don't necessarily have the budget, time, or personnel to investigate. The company gets matched with passionate, hard-working students eager to apply their data science and business knowledge to the real-world. If the project is a success, the company now has the option to rehire the students or propose follow-up projects for new Insight Data Science students. If the project ultimately didn't provide the results that the company was hoping for, then the program acts as an inexpensive way for companies to explore artificial intelligence solutions.

This practicum is geared towards data science and business students but is open to all university backgrounds. Depending on the skill set that each group member brings to the table, the group will have the opportunities to work with the corresponding industry leaders of different fields to solve real-life problems. There are 4 different fields to choose from, namely, information technology, finance and academic research. All groups will be supported by a mentor as a guide and domain expert.

For information technology, we partnered with some of the best tech companies in north America including, Google, Amazon and IBM. There are a lot of exciting projects that our students can be part of, such as virtual personal assistant development, product recommendation, email spam filtering.

For finance analytics, we teamed up with some of the biggest banks here in Canada such as RBC and BMO. The data science and machine learning technology has come to play an integral role in many phases of the financial

ecosystem, the following are some of the current applications of our students can be part of, such as, Portfolio Management, Algorithmic Trading, Fraud Detection.

For insurance analytics, we joined force with well-known insurance companies such as Manulife and Sun life. Undoubtedly, the insurance companies benefit from data science application within the spheres of their great interest, Therefore, we have prepared some exciting data science use cases in the insurance industry for students to participate, such as Insurance Underwriting, Risk Management, Claims Prediction.

For academic research, we collaborated with renowned professors and leading researchers of top Canadian universities such as University of Toronto and University of Waterloo. Here students will have the opportunity to contribute to the technological advancement of the machine learning algorithm development. Some potential research topics include but not limited to Computer Vision, Natural Language Processing, Deep Learning.

FIGURE 11: INSIGHT DATA SCIENCE INDUSTRY PARTNERS