Smith SCHOOL OF BUSINESS Queen's University | Master of Management in Artificial Intelligence

MMAI847
AI Capstone Project Final Report

2021-08-25

Student Name	Student Number
Gopi Balasingam	20253843
Jichen Song	20197262
Felipe Gomes Ferreira	20244867

Filename	Pages	Comments and/or Instructions
MMAI 847 - AI Capstone Project Final.pdf	12	None

Additional Comments:

Capstone Project Final Report

Introduction

Public Yield Capital (PYC) is a convenient partner that matches retail investors and institutions interested in raising capital. Retail investors are important market participants. They behave differently from their accredited counterparts and one another, so the key is attracting the right investors who will participate in the investment opportunities. PYC goal is to provide the tools needed to support valuation, liquidity, and raise capital via investor marketing and communications tactics.

Project Overview

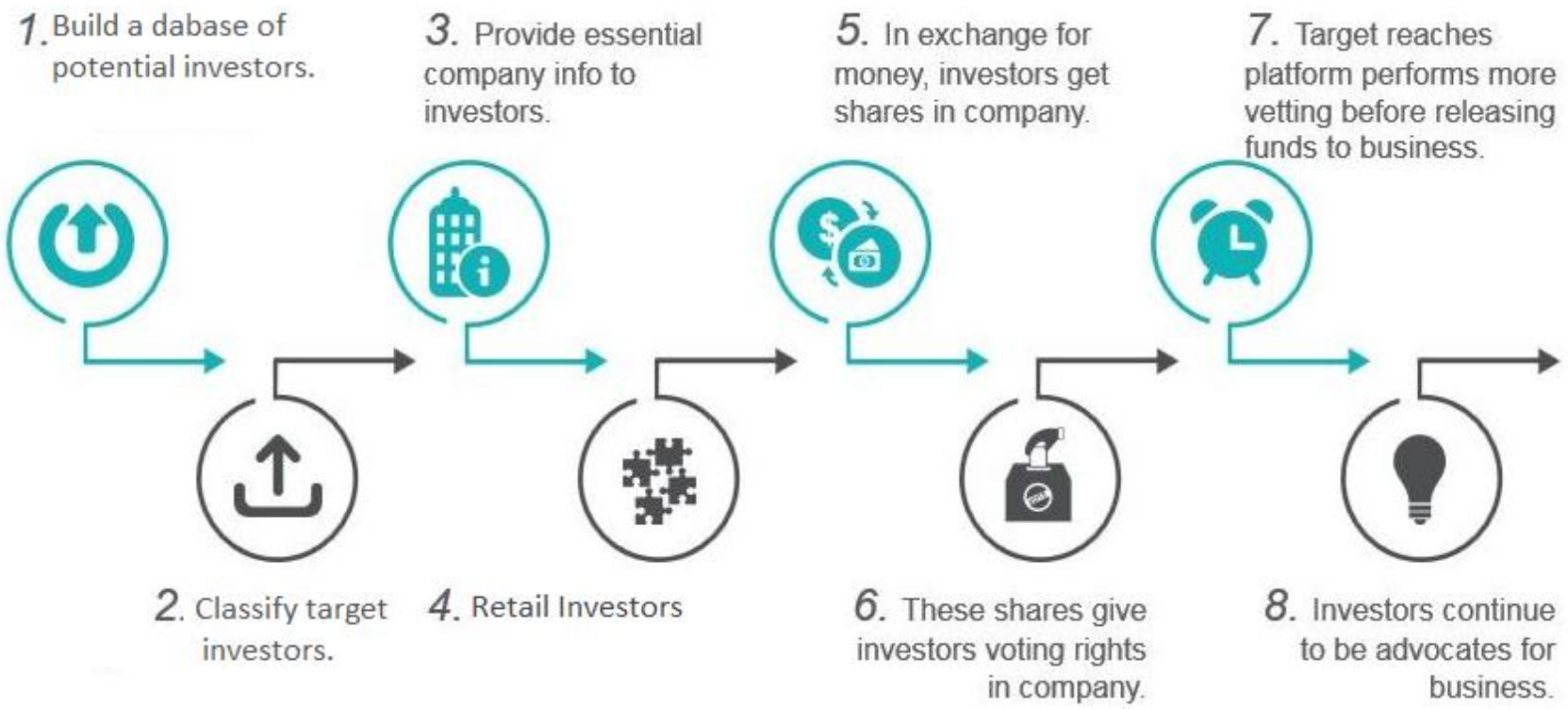
Regulation A+ (Reg A+) allows the public to invest in private companies. Startup companies who are entering the growth stage often owe a large part of their success to their early adopters. By pursuing a Reg A+ offering, a company is inviting its users to share in that success by offering customers a financial stake rewards early adopters for the role they played in the company's growth. Regulation A can be a fantastic fit for a start-up if executed correctly: it allows companies to raise capital from institutional and retail investors, and the ability to raise up to $75 million. Marketing under the guidelines of Regulation A requires careful planning, highly effective advertising methods, and 360-degree coverage. PYC has requested the MMAI capstone project team develop a solution to facilitate Reg A+ investments.

Project Scope

The project will require developing and building AI ecosystem to:

- Build a database of qualified retail investors.
- Develop machine learning classification and clustering models to predict target investors and segment retail investors.
- Develop a deep learning model to predict marketing successes based on investments.

Reg A+ Investment Overview



Project Deliverables

The PYC team provided datasets from a successful investment campaign for Exploratory Data Analysis (EDA) to analyze the dataset and summarize the main characteristics of the dataset using visual methods. EDA helps in finding uncover the different relationship between variables, extract important variables, detect outliers, and maximize insight into a dataset to test underlying assumptions.

Data Collection, Preprocessing, and EDA

To better organize our datasets to be preprocessed and analyzed with the usage of Python programming language, we created csv versions for each data file we received from the PYC team. Initially, the datasets were messy and raw with many duplicated rows, missing data and highly correlated features. In our preprocessing steps, on each dataset, we first removed the less impactful rows which were full of missing values, then we dropped the less useful highly correlated features and then removed duplicated rows. After this step, we merged the datasets of the Gage Cannabis Reg A+ file into a single dataset that contains mostly relevant features.

Figure 1 shows how the value subscribed of the investors are distributed with respect to their country and their status, where the vertical lines represent the confidence interval where 95% of the value subscribed lies in.

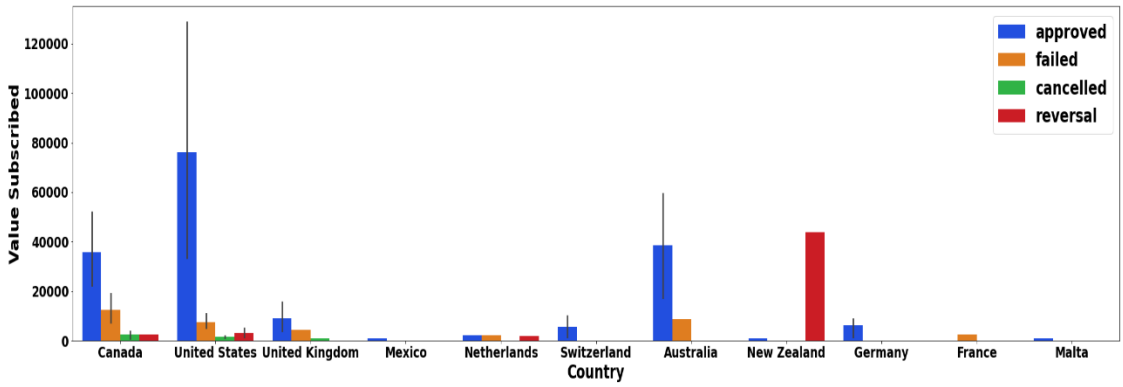


Figure 1. Value subscribed distributed by countries

In our EDA we saw that United States and Canada are the countries that have the biggest number of investors where a higher fraction of them were approved. In response to that, we did a deeper exploration of the distribution of investors around the provinces of each of these countries. For the USA, we saw that although New York state had a smaller number of investors applicants compared to Michigan, California, Florida and Texas, the average of value subscribed of the applications from New York were significantly higher than the national average of value subscribed. Figure 2 show the distribution of the value subscribed of all other states of the USA where we removed New York state because it has an average value subscribed of 1147334, which is far above the range of values of the other states.

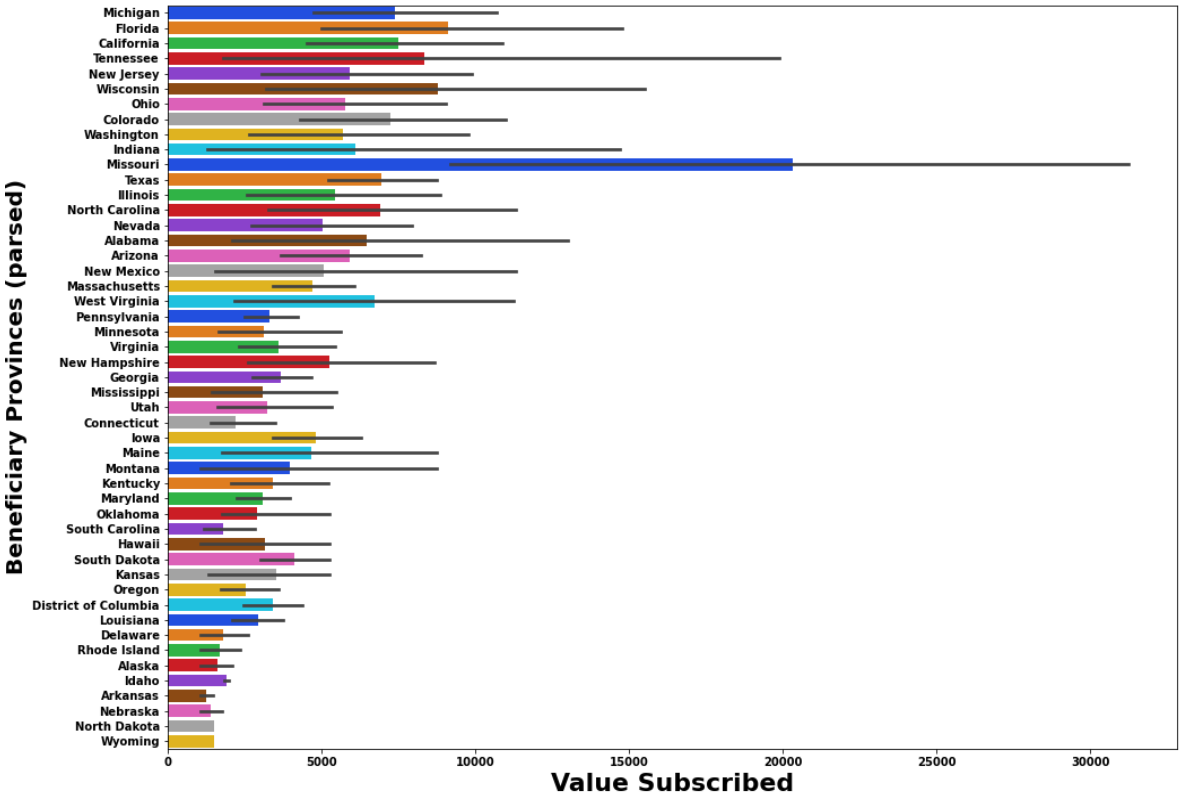


Figure 2. Value subscribed distribution around the USA states other than NY.

For Canada, we saw that Ontario province have the biggest number of investor applications, but the highest ranges of value subscribed were from British Columbia. Figure 3 illustrates the

comparison of the value subscribed from Canada over its provinces including their final status. There were 107 investors whose beneficiary province was Ontario, whereas British Columbia, Québec and Alberta had respectively 10, 10 and 5. More data could provide clearer results.

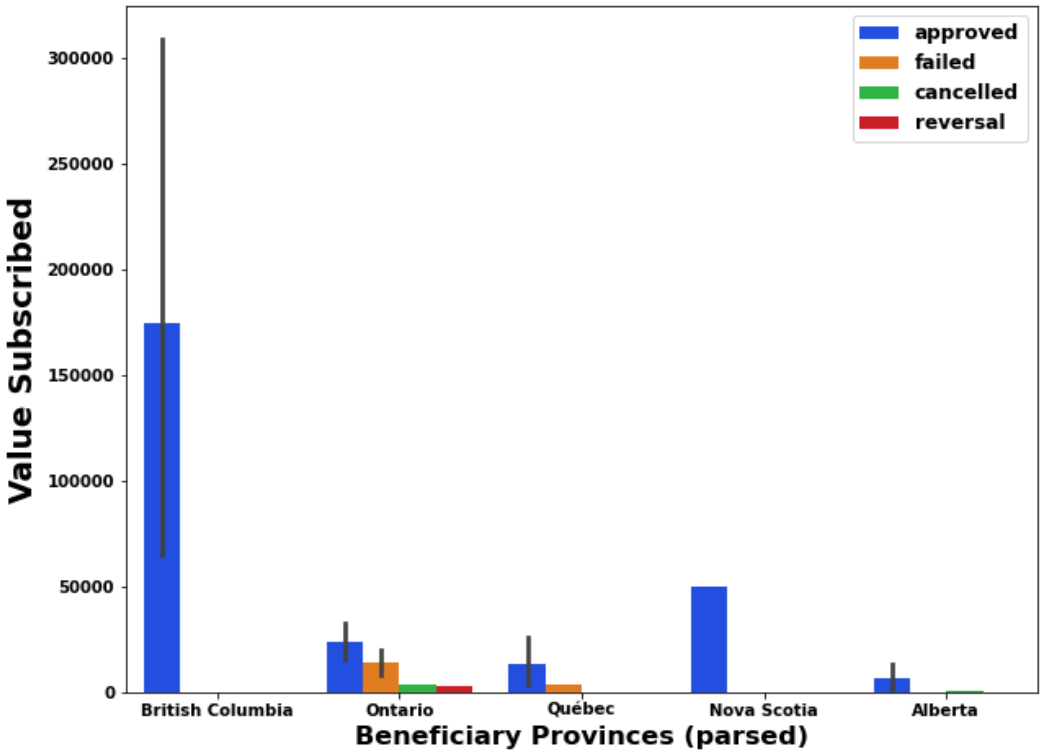


Figure 3. Comparison of value subscribed around Canadian provinces.

Conclusively, with the data we received from PYC, our EDA showed that, for US investors, NY can have the biggest impact on the level of value subscribed of its investors, whereas British Columbia can be expected to have a bigger number of positive Reg A+ investors that are interested in big investments in the case of Canada.

Data Clustering

We performed data clustering for Gage Cannabis Reg A+ transaction dataset. The preprocessed dataset has three features, including value subscribed, beneficiary state, and beneficiary country. The descriptive statistics of the features is shown below (Table 1):

Table 1. Descriptive statistics of the features in Cannabis dataset

	count	mean	std	min	25%	50%	75%	max
Value Subscribed	1174.0	43340.046848	605355.929602	0.0	1085.0	2100.0	5250.0	15625001.0

	count	unique	top	freq
Beneficiary State	1161	70	Michigan	210
Beneficiary Country	1168	17	United States	1019

Agglomerative Hierarchical algorithm was used for the data clustering. The distance metrics used for the numerical features were Euclidean, and the distance metrics used for the categorical features were Hamming. The linkage function for the clustering was Ward linkage. The number of clusters was 4, and the hierarchical result is shown in Figure 4. The descriptive statistics of each cluster is presented in Table 2.

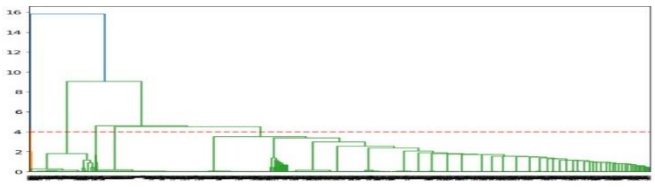


Figure 4. The hierarchical algorithm results

Table 2. Descriptive statistics of each clustering (K=4)

Cluster	Count	Mean	Top State	Top Country
1	3	$ 11,108,334	New York	United States
2	121	$ 25978	Ontario	Canada
3	3	$ 2,666,667	New York	Cayman Islands
4	1047	$ 6125	Michigan	United States

The practice of data clustering divided PYC's Gage Cannabis Reg A+ customers into four different groups that reflect their similarity based on demographic factors. The clustering results allow PYC address each customer in the most effective way to maximize their values to the business. For example, since the cluster 4 has the largest count numbers, PYC may focus on retail investors who are lived in Michigan, US.

Challenges

The topic of strategic data collection was underestimated and PYC is facing data collection challenges in a few specific areas. These areas include:

- Collecting qualitative data about investors
- Converting these qualitative data into quantitative data
- Collecting accurate data about investor behavior and investment preferences
- Verifying that the data collected is clean, standardized, and accurate

Overall, data collection needed for this project was identified to be more sophisticated, emphasizing more on quality, rather than quantity. The main two barriers to solving these problems are:

1. Lack of PYC team bandwidth to identify relevant data sources
2. Lack of capital funding for data acquisition

Solution

Upon presenting PYC management team with our initial EDA results, it was decided that data acquisition is a time consuming and costly process. PYC sees value in the EDA solution and will leverage it to develop meaningful insights for their Reg A+ investment service. Therefore, we will pivot to productizing the EDA solution which will provide PYC with the following capabilities:

- Insights on marketing data to increase investor participation
- Strengthening investor bonds by understanding investment patterns
- Enriching service offerings with new knowledge from analytical insights (i.e. drilling down to individual investment philosophies and preferences)
- Reducing costs by eliminating inefficiencies and human bias

Data Science Practice

The fundamental business aim of data science remains focused on discovering meaningful patterns and producing valuable insights from data as it expands and adds additional "instruments" over time. The data science landscape is composed of several interconnected fields that use a variety of methodologies and technologies So, our recommendation for PYC is

fully embracing data and ramping up their analytics strategy by creating data science team. At a minimum, here are some key roles PYC must consider when building a data science practice:

Type	Function
Project Manager	• Establishes and governs methods and standards of project delivery and tracks project progress • Governs change management and escalation processes
Data Engineer	• Establishes a technology stack for the project • Development of the Reg A+ investment solution
Data Scientist	• Identify data collection tactics • Drive analytics and derive insights • Measure effectiveness of data driven decisions

As data science practices increase in the market, in order to continue growing its business, it will be useful for PYC to move to a cloud system (AWS, Azure or GCP) in the future because it's faster, more efficient and cheaper compared to staying on-premises when dealing with big data challenges. Besides reducing the need for businesses to purchase equipment and build out and operate data centers, cloud computing also facilitates networking collaboration, increase mobility to work from anywhere, provide flexible pay options and is responsible for the security and maintenance of the company's databases. This investment will be highly recommended after PYC build its data specialist team to work with scalable data projects that will be highly impactful for goals of the company.

Data Science Process

The goal of data science is to construct the means for extracting business-focused insights from data. This requires an understanding of how value and information flows in a business, and the ability to use that understanding to identify business opportunities. It is our recommendation that the PYC data science team seek to identify key data assets that can be turned into data pipelines that feed maintainable tools and solutions by the following process:

1. Data Collection and Preprocessing
    - Crawl investor directory listing website and build database of qualified investors.
    - Define ground truth (create labeling documentation).
    - Build data ingestion pipeline and validate quality of data.
    - Label data and ensure ground truth is well-defined.
2. Data Classification and Clustering
    - Supervised classification models (e.g., Random Forest and Logistic Regression) will be trained by the labeled investors dataset to predict target investors.
    - Unsupervised clustering models (e.g., K-means and Hierarchical) will be trained by the unlabeled investors dataset to segment retail investors. The clusters will be interpreted by exemplars and relatively important plots to provide information about the personas of retail investors.
3. Train a Deep Learning model to predict retail investor conversion rate from classification data
    - Use the investor profile obtained from the clustering analysis to combine with the historical time series marketing campaign data to predict which investments are better to each investor profile in such a way that improves the marketing campaign goals.
    - Perform robust training on a large enough dataset to develop a Reg A+ target retail investor model that will provide preliminary insights about the relationship between the marketing campaign and the number of Reg A+ investors.
    - Develop deep learning model(s) to predict Reg A+ investments with respect to advertising campaigns and determine optimal budget allocation for different channels of advertising investment.
4. Model deployment
    - Expose model via a REST API.
    - Deploy new models to a marketing campaign to validate performance.
    - Monitor live data and model prediction distributions.

Operations and model maintenance

In most business problems, Machine Learning Operations (MLOps) require best practices to implement the simplest way to standardize the development, deployment, and maintenance of the ML model via automation. We recommend that PYC adopts the Cross-Industry Standard Process for Data Mining (CRISP-DM) which provides a model for conceptualizing the continuous ML lifecycle. This process begins with the identification of measurable business value, as well as planning for how the solution will be maintained over time. With the end goal in mind, PYC data science team will have to evaluate, collect, and prepare data for training. Once a Machine Learning (ML) model has been trained and tested, it will require optimization to ensure it meets business acceptance criteria before deploying and monitoring in production. This involves a stronger collaboration between the Data Engineer, Data Scientist and the ML Engineer that goes from building clean and organized data sets to feed an operationalizable ML pipeline which can fine tune and retrain the ML model with the new data in an efficient and scalable way. The PYC data science team will evaluate and assess data pipeline, ML models, and the expected quality of production inferences on datasets which is a minimum requirement to successfully adopt a MLOps framework. As such, we recommend the following criteria for the MLOps process:

- Determine any changes can affect the model prediction in unexpected ways
- Periodically fine-tune and retrain model to prevent model staleness and evaluate the data from both a quantitative and qualitative perspective
- If there is a transfer in model ownership, educate the new team members

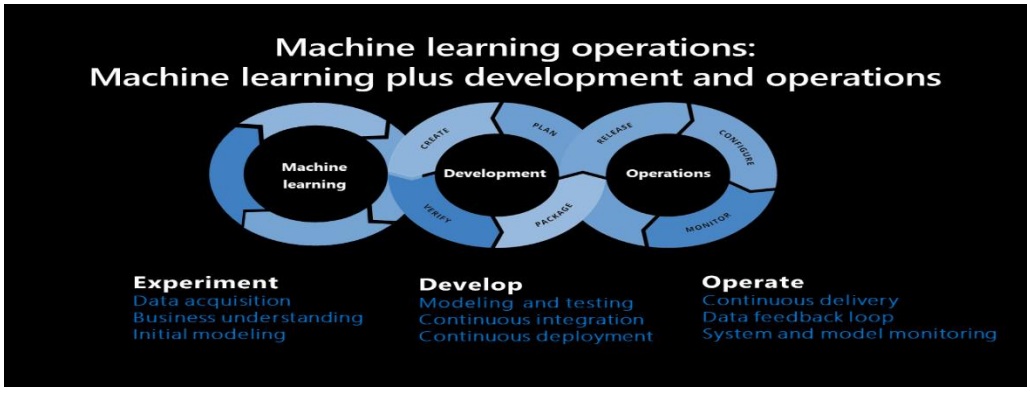


Figure 5. Machine Learning plus development operations.

References

1. Github Repository

   https://github.com/felipeagferreira/Find-RegA-Investors