**INFO 442 - Project Scope**

Tapasya Sharma, Quan Ta, Steven Nie, Kayla Savage

**1. Project Name:** Utilizing Financial Data to Predict Startup Companies Success and Failure

**2. Organization Name:** TKQS Corp

**3. Problem Description:**

**a. What is the problem you are facing?**

Our organization would like to predict small business and start-up success and failure using financial data provided by Intrino. As context, according to the Small Business Administration (SBA), many small businesses' growth and success is vital to the United States economy. Though this is the case, failure rates are high. Specifically, as of 2019, startup failure rates are around 90% and about 21% completely fail by the first year. ([Investopedia](https://www.investopedia.com/articles/personal-finance/040915/how-many-startups-fail-and-why.asp))

**b. Who/what is affected by this problem? (people of certain type, organizations, neighborhoods, environment)**

New business owners, startups and entrepreneurial incubator programs are impacted by the success and the failure of their organizations. Additionally, the economy is impacted by the failure of businesses. An example of this would be, at the start of the COVID-19 pandemic, various businesses closed and as a result impacted the economy heavily.

**c. How many people/organizations/places/etc and how much are they affected? (e.g. mean wait time for surgery, number of students dropping out of school, cost due to tax fraud, etc.)**

As of November 2020, there are 31.7 million small businesses in the United States, considering that small businesses are defined by less than 500 employees; millions of lives are impacted by the growth of these companies. ([Investopedia](https://www.investopedia.com/articles/personal-finance/040915/how-many-startups-fail-and-why.asp))

**d. Why is solving this problem a priority for your organization**

There are various reasons why start-ups may fail for example there could be a lack of resources, little research, wrong market projections, and possibly bad marketing. All of these aspects contribute to the financial success of business. By making this a priority we are contributing valuable conclusions about how businesses can function specifically at the start of establishment.

**4. Goals:**

| **Goals** | **Constraints** |
| --- | --- |
| Finding all start-ups from all available data related to companies | Hard to define which companies are start-ups |
| Predicting the success of these start-ups | The limitation of data related to non-public companies |

**5. Actions**

|  | **Action 1** | **Action 2** | **Action 2** | **Action 4** | **Action 5** |
| --- | --- | --- | --- | --- | --- |
| **Action:** | Acquiring Domain Knowledge | Preparing datasets of companies | Extract insights from the start-ups’ dataset | Predict the success rate of start-ups | Testing the accuracy of the model |
| **Who is executing the action?** | All members of the team contacting ideal specialists | All members of the team | All members of the team | All members of the team | All members of the team |
| **Who/what is the action being taken on?** | Accountants and Financial Economists | Retrieving the data using Intrino API  Cleaning the data | Exploratory analyzing the datasets while looking for key features to identify start-ups and determine their success | Building models to train and test the datasets | The model accuracy and adaptability to real-world data |
| **How often is the decision to take this action made?** | Throughout the project when we have questions about the data but majorly in the beginning while we get used to the data | Once or twice in the project based on the data | Multiple times until find reasonable feature | Multiple times until finding desirable results | Once or twice in the end before final deployment |
| **What channels are/can be used to take this action** | Intrino API | Intrino API | Python | Python | Python |
| **Other useful information about the action** | NA | NA | NA | NA | NA |

**6. Data**

For this project we will collect data via Intrinio API, one of the largest databases for US and global company data. Intrinio provides many functions in its API but we will mainly use historical data tag and some search company functions to get a list of companies we are interested in. We will also write out code needed to collect data via the API, based on the project goal we proposed with all the appropriate data points. Lastly we will also need to learn and organize the data in a way that can be used for building a model.

Some examples of the data tags that we would be using are:

1. **Basic Info** - Dividend per share, Number of employees, Country
2. **Calculation and Metrics** - Accounts Payable Turnover, Capital Expenditures, Cost of Revenue to Revenue,
3. **Stock Price Data** - IPO offering price, delisted T/S, current price, 52 week high, 52 week low, ask price

**A. What data sources do you have internally?**

NA - We would be using external data.

**B. What data can you get from external, private or public sources?**

|  | **Data Source 1** | **Data Source 2** |
| --- | --- | --- |
| **Name** | Intrinio API | SEC.gov |
| **What does it contain?** | Historical financial data on US company and securities | SEC filing files |
| **What level of granularity?** | All SEC Filing and financial report data point of companies and some commonly used metrics derived from these data | Original file source for any financial report and filling |
| **How frequently is it**  **collected/updated after**  **it’s captured?** | Only captured once since we are only working with historical data | Only captured once since we are only working with historical data, but may use it regularly to check the accuracy of the data |
| **Does it have unique**  **identifiers that can be**  **linked to other data**  **sources?** | Ticker symbol like AAPL for apple Inc.and company name | Ticker symbol like AAPL for apple Inc.and company name |
| **Who’s the internal**  **owner of the data? Where will it be stored?** | It will be stored on our computer probably as a csv file for easy access by python | Most likely won’t be stored since it will only be used occasionally to check for data accuracy |

**C. In an ideal world, is there additional data you would want to get/gather that would be relevant to his problem?​(surveys, CCTV, phone records, DNA, different frequency or granularity for currently available data, etc)**

Private financial data from a small startup company, it would be immensely useful if anyone can get their hands on that but in practice it is virtually impossible for an outsider to obtain.

**D. ​What analysis will the existing data not support? ​For example, if we don’t have access to outcomes for students in wich case any analysis predicting the outcomes will not be feasible until the outcome data (or a reasonable proxy) is collected**

The data that we are currently using is only limited to the financial aspect of predicting if a startup or a company will succeed or fail. We do not have data relating to aspects such as how fast their growth was, how many years did it take, did the growth being stagnant.

**7. Analysis**

|  | **Analysis 1:**  **If the IPO startup company will succeed or failed** |
| --- | --- |
| **Analysis type** | True/False classification |
| **Purpose** | Provide an alternative AL model to analyze if company will likely to succeed |
| **Which action will this analysis inform?** | Investment decision |
| **How will you validate this analysis using**  **existing data?What methodology and what**  **metrics will you use? How will you compare against existing baselines?** | We will split the dataset into training and testing data, but for a real world validation it would not be possible as company takes years to take shape |
| **What**  **limitations will**  **this analysis**  **have, either**  **based on**  **available data**  **or choice of**  **methodology?** | This prediction will never be really accurate since it strictly rely on the financial report data reported by the companies itself which can be manipulated or outright forged. Additionally many unexpected factor will impact a company that are beyond the scope of their financial reporting, like an pandemic will severely affect an airline whom won’t be able to predict the pandemic itself |

The team would define what counts as a successful start-up and what is not. It will involve the status of the stock on the main stock exchange. For example, a delisted company can mostly be interpreted as a failed company, but we will also observe its stock price to break it down further. If the stock price failed to hold its offering price a few months after IPO it can also be interpreted as the company failing to live up to the expectation and most likely on a downward spiral. The team will also find business data points that should be used in the model for prediction ie. the metrics and calculation which would be useful and relevant in a prediction model. Lastly, we would come to define some explanation of the chosen data point and their real world business case usage since it does have some interesting real world application.

**8.** **Ethical Considerations**

| **Privacy** | There are no major ethical concerns for our project Intrino API sources their data from publicly available SEC filings. |
| --- | --- |
| **Transparency** | There are no stakeholders that need to be aware of this project. It is fully meant to explore various methodologies for assisting and adjusting small business methods. |
| **Discrimination/Equity** | This model would likely contain data from financial reports made from 2020. Because of the impact of COVID-19 on various companies, our model will show a heighted failure rate. (Dependent on split dataset)  Data points will have to be selected carefully as we sort through companies with different business models and tools such as inventory models versus software distribution. |
| **Social License** | This model would not be super useful with the reporting of companies all over America. We hope that after deployment, this tool can be used to improve business models and give insight on how machine learning can provide insight to company financial data. |
| **Accountability** | TKQS is responsible for the implementation and use of this model for evaluating how small businesses succeed or fail over a time. |
| **Other** | As a company, we can not decide how information is interpreted. |

**9. What field trial or randomized controlled trial can you design to validate the project in the field? The outcomes you will measure should match your goals. Define the population in which the model will be tested. Define the duration of the trial. Specify the baseline. You should measure the impact in different population subgroups**

Our project is based on predicting if a startup will succeed or fail which we would be analyzing based on the financial data of companies in the past that have succeeded or failed. A field trial in our case would pertain to selecting an ideal model, seeing if the model works successfully with new data and deployment -

1. Identify the type of model we want to use - This would involve experimenting with ML models (mainly classification models for our binary target value) and understanding which works the best.
2. Cross Validation - We would also be identifying training and testing data and check for accuracy through the testing data. Accuracy would be further increased by a possible 5 or 10 fold cross validation method. A baseline accuracy that we aim for is at minimum of 70%.
3. Deployment - Another way to test is if the project would work with real life data outside of our set testing and training data and see if the model can be used by others as well. Here, others refer to populations such as accountants and economists who could potentially use our model if they value it. Based on financial decisions, this model could identify the “next big thing” in startups for people who understand the financial and the technical space.

**10. Who are the external organizations and internal departments that will need to be involved? (Typically, data science projects need involvement from data owners, IT infrastructure owners, problem owner, analytics people)**

| **Organization/Department** | **Description of desired involvement** | **Name/role of counterpart** |
| --- | --- | --- |
| **Accounting Teams** | A lot of the data we are using for this project includes terminologies that accountants are familiar with and use on a daily basis. We need external input to understand and interpret our data better. | Accounting Students or Professor |
| **Stock Market groups** | Since our data is based on the financial evaluation of companies, we need input from financial economists that understand the stock market, how it fluctuates, what influences it and what we should be looking for in our data. We could also look for this knowledge from existing YouTube channels and/or information already on the internet from financial economists. | Financial Economists |