Independent Study – Midterm Report

1. Introduction

Machine learning can be used to classify and cluster data sets to potentially identify complex trends and relationships within the data which are not immediately noticeable through other routes of analysis. This is particularly true for data sets with many variables, or features, per item, as the relationship may be multivariate and non-linear. In this case, machine learning may provide more insights by allowing for holistic analysis. For instance, the field of computer vision which attempts to identify objects within data, Although most major applications of machine learning focus mainly on analysis of purely quantitative data (Bishop, 2006), one can both classify and cluster qualitative data as well based on conceptual understandings (Fisher, 2018).

However, the results of machine learning algorithms must be statistically analyzed to test the validity of the classification/clustering. The most fundamental way of testing the algorithm is to understand the general accuracy, how well the algorithm's prediction of a class or cluster matches the expected, pre-solved answer. However, there are other tests which measure the sensitivity of the algorithm, how well it responds to changes in the data set, or its precision, how consistently it classifies/clusters (Sokolova & Lapalme, 2009). As such, it is clear that statistics and machine learning should be used in tandem to ensure that the insights machine learning gives are not flawed.

Economists have also been harnessing the power of machine learning and statistics to test and investigate hypotheses which were previously simply assumed to be true. Within this intersection of economics and computer science, I hope to introduce a rigorized analytical approach to the search fund process, wherein an individual entrepreneur, or search fund principal, seeks to acquire and grow a company, with the potential of selling it at a significant profit after 5-10 years of leadership. The search fund model is continually being refined and renewed, with the idea only being coined in the mid-1990s. In 2016, Stanford's Center for Entrepreneurial Studies published a general search fund survey finding that, in comparison to 2009 search funds, recently successful search funds spent increasingly more time searching for possible companies in various industries and less time personally contacting acquisition targets and performing due diligence research. However, 27% of all search funds historically fold without making an acquisition, despite reviewing numerous prospects (Kelly et al., 2016). This suggests that if one could quickly survey possible companies without having to perform highly in-depth research, there could be a heightened chance of finding, negotiating, and hopefully acquiring a good target business without excessive use of funds. Currently, however, most economic issues are modeled with very large numerical datasets and relatively transparent, reliable information, with previously proposed algorithms for ranking companies for investment focused more-so on quantitative firm data such as amount of funding or number of employees (Athey, 2018). However, for economic models which include more social definitions, such as a general evaluation of a company's potential growth, qualitative data such as brand image is necessarily included (Stanford, 2017), thus making it clear that one ought to analyze mixed data models, wherein features are both quantitative and qualitative, instead of purely quantitative data.

The key challenge faced during the search phase is information collection and analysis. Per Stanford's 2018 Selected Observations Study (2018) of search funds, search fund principals spend increasingly more time screening potential companies before pursuing. This may be due to increased competition in the space as well as a need for more judicious selection of a stable business model and industry due to the current economic trends. Regardless, however, it is clearly increasingly important to be able to quickly and accurately assess potential companies in order to efficiently proceed to contacting and hopefully acquiring a target company. Because target businesses are privately owned, however, it is difficult to obtain critical information which can factor heavily into whether the company may remain functional post-acquisition. For instance, the search fund I work with aims for a business with a stable recurring revenue model and an EBITDA between $1-3 million. However, estimated revenue and capital structure information are not always available, nor are they necessarily accurate. Unlike traditional private equity, there is a distinct lack of information regarding target search fund companies, where preliminary information is typically only what is gathered through online presence. Thus, the current search fund screening strategy involves a generalized intuition regarding whether the business is enduringly profitable, within the target range, and possesses room for growth. While some characteristics distinctly raise the likelihood of being chosen for further target, others hold ambiguity and are evaluated in case by case situations. For example, when considering regionality, manufacturing companies tend to stay regional because cost is high for them to expend. In contrast, it’s much easier software companies to go national or global, but because all software companies can do that, the market is also competitive. Thus, the value of regionality alone is not sufficient to drive decision making but has to be evaluated with a combination of other features.

As such, in this independent study I aim to investigate whether modern machine learning algorithms can be used against such a particular mixed data set which contains numerical and categorical data. Since algorithm choices can and should be tailored to best model the understanding of the model it is meant to find connections within, I will likely consider multiple algorithms. Furthermore, I will study whether one could find a quantitative weighting for various features associated with a company and whether various relationships between these features can be used to predict if the company should be further investigated as a possible acquisition target through use of supervised machine learning algorithms and various statistical analyses. If possible, such established weightings and/or relationship analysis would better shape the understanding of the complex and novel characteristics suited for an acquisition, and augment search fund strategy by optimizing the time and types of research needed during the search phase.

2. Data Collection/Formulation

To gain intuition regarding the presumed standards for potential acquisition targets, I worked with the search fund principal and main analyst to collate the most important categories to consider regarding business model characteristics and general audience on an industry and company level, as well as some general intuitions regarding which categories were more important. These observations are summarized below:

Table 1. Industry and Company-Specific Criteria for Potential Search Fund Acquisitions

Industry-Level Criteria	
Criteria	Importance
Recurring Revenue Model	The general revenue model of an industry or industry niche gives an inexperienced CEO greater visibility into the expected revenue pipeline, which facilitates budgeting, long-term planning and valuation. This also allows an inexperienced CEO to take the time to truly understand the business following acquisition rather than having to refresh revenue pipeline immediately.
Healthy and Sustainable Margins	High EBITDA margins (15%+), like growing industries, give inexperienced CEOs cushioning in the event of economic down-turns or other bumps in the road. Also, high margins provide a cushion for increased investment in the business (e.g., hiring additional sales people, upgrading IT systems) if the seller had under-invested.
Stable Cash Flows	Stable cash flows will ensure that the company will be able to service its debt. In part, stability is achieved through a recurring revenue model and

	predictable cost structure. In part, positive cash flows are achieved through low capex.
Growing Industry	A growing industry provides a tail-wind. In theory, a rising tide raises all boats, a company in that industry will experience growth without the need to make drastic changes to the business, increase market share, develop new service lines, etc. This is a more favorable competitive environment, especially for an inexperienced CEO. By contrast, declining industries often deteriorate into a zero-sum game, increasing competitive pressure as companies compete primarily on price. Kessler and Ellis in "Search Funds - Death and the Afterlife" list low/negative growth as "Theme One" among unsuccessful search fund acquisitions.
Fragmented Industry	Fragmented industries (1) are less likely to have a dominant player that distorts the competitive environment (e.g., through lowering prices in order to increase market share), which makes for a favorable competitive environment, (2) provide more targets for acquisitions, (3) valuations will be more reasonable because there will be less strategic buyers, and (4) provide more avenues for growth through product/service or geographic expansion.
Straightforward Operations	An industry with straightforward operations will allow an inexperienced CEO to focus on increasing value immediately, rather than having to spend time researching company and industry dynamics. Kessler and Ellis in "Search Funds - Death and the Afterlife" list complex operations as "Theme Two" among unsuccessful search fund acquisitions.
Low Exogenous Risk	Exogenous Risk (e.g., technology change, regulatory, litigation, environmental, commodity exposure, seasonality/cyclicality) could adversely impact the entire industry and majorly disrupt the traditional business model. This may result in a growth industry turning negative; barriers to entry could be eliminated; margins could disappear; a major substitute could be introduced. An example would be how peer-to-peer ridesharing has disrupted the taxi industry by leveraging common drivers, a large resource, to overcome specialized taxi driver knowledge.
High Barriers to Entry	If threat of new entrants is high, there will be pressure on (1) market share, (2) price, (3) costs and (4) rate of investments.
Low Intensity of Rivalry	Highly competitive industries tend to lead to price competitivity which erodes margins and makes sustainable growth more difficult.
Low Customer Power	Powerful customers will capture more value by (1) forcing prices down, (2) demanding better quality/service, (3) playing companies off one another.
Low Supplier Power	Powerful suppliers can capture more value by (1) charging higher prices, (2) limiting quality/service or (3) shifting costs away. Ex: Microsoft eroded the margins of computer makers by increasing the price of Windows operating system.
High Switching Costs	High switching costs lock customers in and enable you to incrementally raise prices every year without worrying that the customers will find better alternatives with similar characteristics or at similar price points.

Few/No Substitutes	Too many substitutes can result in a price ceiling (so as to avoid consumers pursuing substitutes) and customers will thus have high bargaining power.
Company-Level Criteria	
Criteria	Importance
Appropriate Size	If too large, the entry multiple will be higher and operations are more likely to be complex, among other things. If too small, there is less money to invest in growth, smaller room for error, and it is less likely that the company will have a track record of growth, among other things.
Track Record of Growth and Profitability	Turn-arounds are complicated and are often unsuccessful. A track record of growth and profitability provides a stable base for an inexperienced CEO to grow incrementally.
Healthy and Sustainable Margins	See industry criteria.
Recurring Revenue Model	See industry criteria.
Stable Cash Flows	See industry criteria.
Low/No Customer Concentration	Losing a large customer can have an out-sized impact on a business if concentration is high. This risk is especially acute if the seller has a strong relationship with the key customer. Additionally, a key customer can exert a lot of power, thus taking value for itself.
Realistic Exit Options	Search fund CEOs aim to significantly grow the business, but also will likely aim to sell or transfer company power in 5-10 years time. Exit options are realistic if there are strategic buyers in the industry or if you experience enough growth to interest PE shops, among other things.
Strong Management Team	A strong management team will make the transition easier for an inexperienced CEO as well as increase likelihood of maintaining important client relationships and institutional knowledge.
Qualified Seller	There is no need to spend time/effort on a seller who is not willing to commit to a transfer of power. Need to find someone who is both willing to sell and willing to sell to an inexperienced CEO.

In this particular study, however, all information will be taken from online company information, usually the company’s professional website as well as social media (Facebook, LinkedIn) offerings; this typically means the company’s financials and general business model is unknown, and thus it is not completely certain where a given company ranks in most categories listed above. Some of these categories are very unlikely to be known given that companies to not tend to advertise their availability for acquisition, and thus the notions of having a strong management team which will stay post-acquisition and a willing seller are not particularly predictable from the onset. Moreover, the company-specific revenue model is not necessarily obvious from the outset, particularly given that many smaller companies utilize a mixed model of offering projects/products, which tend to be less recurring in orders, and services, which tend to be more recurring. Even within projects/products, we can envision a difference in offering products which are a one-time purchase versus those which are purchased using a regular subscription fee. Further, there is a perceived distinction between repeat business, in which clients fundamentally purchase each service/good once but are very likely to purchase again, and recurring business wherein the service offered is extremely sticky, such as utilities. While theoretically there exists a spectrum between totally non-recurring and totally recurring revenue, it would be very difficult to immediately determine where a company lies solely given public information.

Rather, it would be more reasonable to value each company ordinally from 1 to 5 in each qualitative category, wherein 1 represents a distinct lack of fulfillment and 5 a near total lack of fulfillment. While somewhat arbitrary, this allows for more easy ranking within and between categories. From this, I have currently narrowed down the desired features to be analyzed for a given potential acquisition company to the following: margin levels, cash flow stability, industry growth, industry fragmentation, straightforwardness of operations, possibility of exogenous risk, barriers to entry, stickiness of offerings, minimization of supplier power, minimization of customer power, service to project/product ratio, level of long-term client relationships, minimization of customer concentration, industry niche, and employee estimate. The first 10 features are directly taken from the expert opinion surveying, wherein firms are expected to be better when ranked higher in these values. The next three features (service to project/product ratio, level of long-term client relationships, and minimization of customer concentration) aim to measure how likely it would be that a firm has recurring revenue, which hopefully can be determined by providing more service and/or more long-term client relationships and/or having multiple clients at any time. Such things are likely emphasized in company newsletters and general promotions, and thus are more defined given public information. The industry niche would be a written description of the general function of the company. Finally, the employee estimate, per the search fund principal, is a decent way of estimating revenue amount, with most companies within the ideal range usually ranging from 10-200 employees. The employee estimate is bucketed per LinkedIn estimates from 2-10, 11-50, 51-200, and 201-1000. These would all then be mapped to whether the company was accepted or rejected for further pursuit.

Thus, the characteristics desired for a particular company relate to both its industry niche as well as its own personal characteristics, particularly relative to the general industry. Given that companies are discovered on an individual basis and that the evaluation of industry niche will have to also be somewhat based on the characteristics the companies which fall under its categorization, we see that we can either analyze a dataset of a given company's features on an individual basis or collated with multiple companies in a similar industry position.

The latter collation and analysis requires some way to cluster industry niche descriptions, which we can envision in multiple ways. For a given industry niche description such as "miscellaneous chemical manufacturing", "computer system design services", or "medical equipment and supplies", there would need to be an algorithm collating the similarities between descriptions. Because most machine learning algorithms only accept numeric inputs, the categorical or descriptive text-based features needs to be vectorized using natural language processing methods such as TF-IDF or One-Hot-Encoder. For example, with TF-IDF, distinctive words in data fields such as "Industry Niche" will be converted into an array of vectors. The output of the TF-IDF vectorizer can then be converted to a list before adding back into the original data set and ingest into machine learning models. Alternatively, new machine learning and analytical platforms, such as H2O.ai, with enhanced features in handling categorical data will be investigated. H2O.ai offers advanced categorical encoding features such as Frequency Encoding and Target Mean Encoding which enables us to encode features in respect to frequency of occurrence or means associated with positive outcomes. Risk of overfitting would then be reduced through k-fold cross-validation or computed weighted means cross all training set and even level.

Regarding the former analysis, I will analyze the features and categorize the companies as accepted or rejected for further due diligence. Then, I will ingest the same list of companies along with the collected features into various machine learning models, using the categorized companies for training, and generate predictions. Supervised learning is particularly suited for this data, as its classification is labeled with accept or reject, and it possesses various known features of unknown weight to the classification. Using various algorithms like support vector machines (SVM), k-nearest neighbors (k-NN), or Bayesian classification, one can both find possible relationships between variables as well as note the most important factors by their impact on the final classification (Kotsiantis et al., 2007). I will likely be focusing on algorithms which either classify a given business as accepted or rejected based on its description or place it in a cluster of similar companies to perhaps find unforeseen similarities between feature groups. Per Raudys et al. (1991), the limited sample size of my data may lead to heightened sensitivity to algorithm parameters, such as the number of neighbors in k-NN, and as such should be monitored and analyzed. Techniques as such as Mega Trend diffusion (MTD) or adding Synthetic Attributes can be used to overcome limitations from small dataset. Pre-processing data using such techniques may significantly improve classification accuracy (Ruparel, etc, 2013). Additionally, I will research and hopefully implement algorithms which are optimized for handling smaller sample sizes, such as the one proposed by Hu et al. (2015). In the event of a missing field for a data point, multiple imputation may be used to introduce the notion of uncertainty regarding the value (King et al., 2001).

Afterwards, I will rank the algorithms on their general accuracy on an unlabeled test data set, where the algorithms predict the final classification or clustering of the business based on its features. Moreover, I can further use statistical tests to determine the general precision of the algorithm, such as allowing the algorithm to parse the same business multiple times and analyzing whether the classification varies heavily, a type I error (Dietterich, 1998). It has been shown in comparative studies using the same dataset, that SVM produces lowest Type 1 (false positive) error, Decision Tree produces lowest Type 2 (false negative) error and highest overall classification accuracy (Chen et al., 2014).

Current Bibliography

1. Athey, S. (2017). The impact of machine learning on economics. *Economics of Artificial Intelligence.* University of Chicago Press.

2. Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* New York: Springer.

3. Chen, S., Goo, Y.-J. J., & Shen, Z.-D. (2014). A Hybrid Approach of Stepwise Regression, Logistic Regression, Support Vector Machine, and Decision Tree for Forecasting Fraudulent Financial Statements. *The Scientific World Journal, 2014*, 968712. http://doi.org/10.1155/2014/968712

4. Creswell, J. W., & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.

5. Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7), 1895-1923.

6. Fisher, D. H. (1987). Knowledge acquisition via incremental conceptual clustering. *Machine learning*, *2*(2), 139-172.

7. Hu, Y., Guo, D., Fan, Z., Dong, C., Huang, Q., Xie, S., ... & Xie, Q. (2015). An improved algorithm for imbalanced data and small sample size classification. *Journal of Data Analysis and Information Processing*, 3(03), 27.

8. King, G., Honaker, J., Joseph, A., & Scheve, K. (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American political science review,* 95(1), 49-69.

9. Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160, 3-24.

10. Ruparel, N. H., Shahane, N. M., & Bhamare, D. P. (2013, May). Learning from small data set to build classification model: A survey. In *Proc. IJCA Int. Conf. Recent Trends Eng. Technol.(ICRTET)* (pp. 23-26).

11. Raudys, S. J., & Jain, A. K. (1991). Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (3), 252-264.

12. Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*(4), 427-437.

13. Stanford Graduate School of Business (2017). A Primer on Search Funds.