Selling Ideas

Examining the Relationship Between Startups' Patents and Their Ability to Raise Funds

Jason Knobloch
March 15, 2015

Introduction

Is there a correlation between the claims made in a patent and the investment decisions made by venture capitalists in terms of which startups they choose to back? This is a question that machine learning techniques may help answer given the public availability of patent applications and grant documents and web-based aggregators of startup company information. Understanding the relationship may prove useful to startups pursuing venture capital, venture capital firms seeking to measure potential investment opportunities, and policymakers trying to understand innovation dynamics in local, regional, and national economies.

Background

"Innovation" has been a prominent buzzword in the American business lexicon for over a decade. Startups with new ideas have fundamentally altered many industries and created new ones. This is made possible, in part, by taking advantage of advancements in hardware and software that have lowered to almost nothing the capital requirements for bringing services to market. Startup companies' ideas, their intellectual property, are therefore among their most valuable assets. Patenting these ideas, products, and services is a fundamental method of protecting this intellectual property and maintaining a unique position in the market. Although using patents as a measure of innovation is not without controversy, evidence suggests they are a useful metric when considered carefully.[1]

Venture capitalists seeking to invest in startups face an information gap relative to investors in established firms. Mature companies have measurable revenues, cash flows, assets, and so on that indicate their value. Startups have none of those things as they have, by definition, yet to produce or sell anything. Investors

[1] Gregory Ferenstein, "Patent Director: 'Patent filings do not equal innovation,' US Needs New Measure," *Fast Company*, March 12, 2011, http://www.fastcompany.com/1738089/patent-director-patent-filings-do-not-equal-innovation-us-needs-new-measure; Riita Katila, "Measuring Innovation Performance," *International Journal of Business Performance Management*, vol. 2: 180-193; Bronwyn Hall, "Using patent data as indicators," presentation to the European Meeting on Applied Evolutionary Economics, June 2013, http://eml.berkeley.edu/~bhhall/papers/BHH13_using_patent_data.pdf.

instead look to the management team's ability to execute, the size of the market the company is targeting, and the novelty and latent demand of the product or service the company proposes to create as the key factors for funding decisions.[2]

With this in mind, the intent of this research is to utilize the text of the patents to develop a 'uniqueness' value as a metric of the novelty of a product. This value, combined with the industry sector market capitalization as a metric of the size of the pertinent market, will then be compared with a response, namely the startup funding the holder of the patent received.

Data Acquisition and Pre-Processing

US Patent and Trademark Office (USPTO) patent grant data was obtained from the bulk download website hosted by Google via a simple web scraper. A zipped file of patents is released weekly, and full text patent grants are available from 1976 to the present day. Files containing patents prior to 2001 are in basic text (.txt) format, while those from 2001 on are in extensible markup language (.xml). The intent had originally been to acquire all of the available xml's, but as the size of each unzipped file is between 400MB and 700MB, space limitations limited the data available for this project to the patents released between January 3, 2012, and October 7, 2014.

Parsing the patent xml's was a challenge. Python xml parsing modules require the file to be 'wrapped' with some variation of <root> and </end> tags, tags the patent xml's do not have. The first action taken was to rewrite the xml files by adding the tags <patents> and </patents> to the beginning and the end of each file and skipping lines in the body of the file that interrupted parsing. Then, utilizing the C implementation of the ElementTree module,[3] the following data was extracted:

1. The assigned patent number
2. The invention title
3. The date of the patent grant
4. The date the application was made
5. The main US government classification of the patent
6. The number of references cited by the patent applicant and patent examiners
7. The holder(s) of the patent
8. The city, state, and country of the patent holder(s)
9. The number of applicants

[2] Ben McClure, "How Venture Capitalists Make Investment Choices," Investopedia, http://www.investopedia.com/articles/financial-theory/11/how-venture-capitalists-make-investment-choices.asp, accessed March 14, 2015.

[3] The C implementation of the ElementTree module is well documented, performs well, and has better XPATH support than some other xml parsing modules such as Beautiful Soup.

10. The full text of the claims made in the patent, i.e., the patentable product or idea

Several of the fields required cleaning, either as a result of the extraction process or due to the messiness of the provided inputs. A complicating factor was that the original xml's are encoded in Universal Character Set and Transformation Format – 8-bit (utf-8). Python 2.7, by default, encodes to American Standard Code for Information Interchange (ASCII), and occasionally displays quirky behavior when working with utf-8 strings.[4] This was accounted for while removing extraneous characters introduced during data extraction.

Another issue was dealing with misspellings in the provided data. Just to provide one example, the official name of the auto company BMW, Bayerische Motoren Werke Aktiengesellschaft, was spelled sixteen different ways. No good solution was found to deal with this issue, although this is not believed to have affected the outcome of the project.

To facilitate the gathering of industry sector market caps, it was useful to append the title of the patents' main classifications to the dataset. Although formatting of the main classification numbers was inconsistent, an adequate solution was found that enabled the cross-referencing of the first three characters of the main classification number with the classification section headings used by USPTO.

Information on startup companies and their funding rounds was gathered from Crunchbase, a site that catalogs startup activity, through utilizing the site's API and the python module pycrunchbase. The API calls searched Crunchbase using the cleaned assignee organization names from the patents dataset, thereby ensuring matching key values between the patent data and the funding data. The funding data consisted of:

1. Company name
2. Date of funding round
3. Funding round amount in US Dollars
4. Funding type (angel, seed, venture, etc.)
5. Series, where applicable (Series A, Series B, etc.)
6. Universally unique identifier (uuid) assigned by Crunchbase

Industry sector market caps, used as an indicator of market size, were based on industry indices provided by Yahoo! Finance. Patent main classification titles were matched with the particular industry sectors that seemed most applicable and the 2014 market capitalization values were coded by hand. A more sophisticated and repeatable approach was considered, but research indicates

[4] See "Overcoming frustration: Correctly using unicode in python2," https://pythonhosted.org/kitchen/unicode-frustrations.html, accessed March 15, 2015.

that matching patents to specific industries is a long-standing and poorly addressed problem.[5]

The key variable that was derived from existing data was the ‘uniqueness’ value for the text of the patent. The text of all patent claims was combined into a single corpus. The sum of term frequency-inverse document frequency (TF-IDF) values of each patent claim was then calculated and multiplied by the number of non-stop-word tokens in the claim, resulting in the uniqueness value.

$$\textit{Uniqueness} = \textit{Sum(TF-IDF Values)} \times \textit{Number of Tokens}$$

For example, a nine-word claim describing a pair of sunglasses received a uniqueness value of 7.32, whereas a several-thousand word claim describing the systems and methods for network virtualization received a uniqueness value of 2605.59. The range of uniqueness values can be described as:

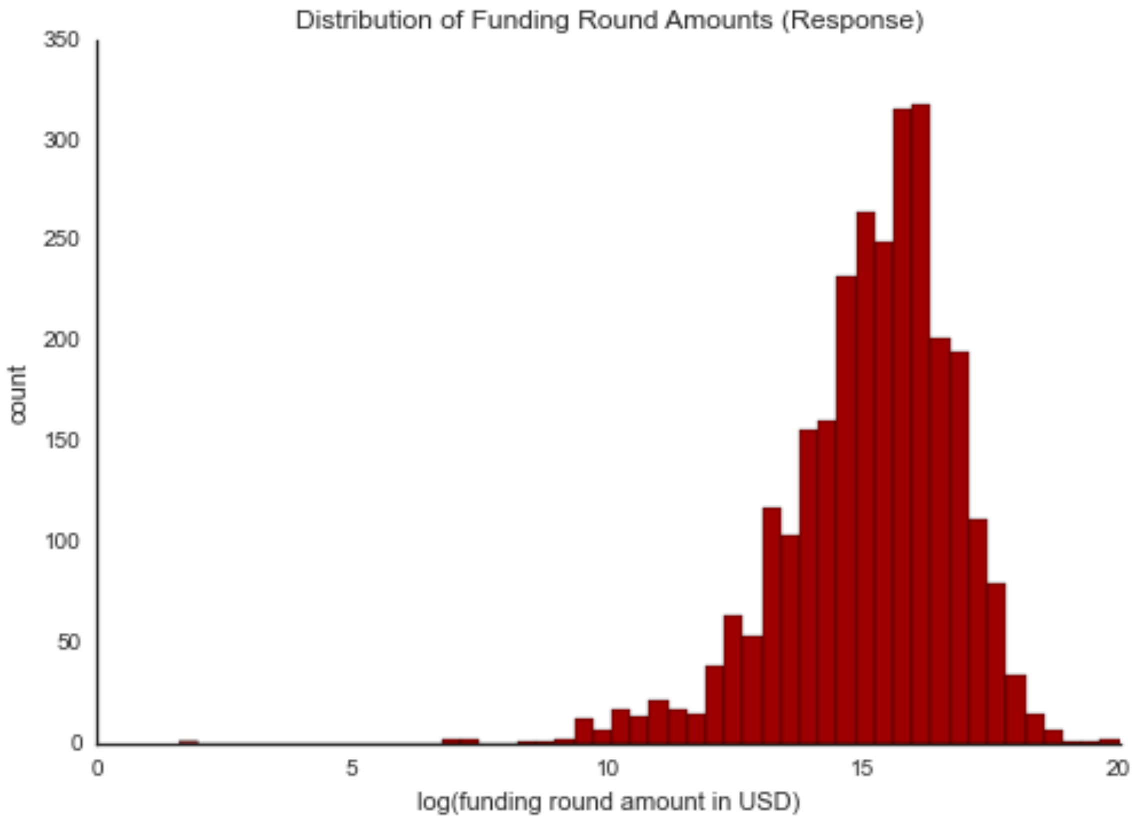
Uniqueness Values	
Mean	734.498034
Standard Deviation	569.970717
Minimum	7.321309
25%	419.510578
50%	616.545718
75%	924.421893
Maximum	9790.160784

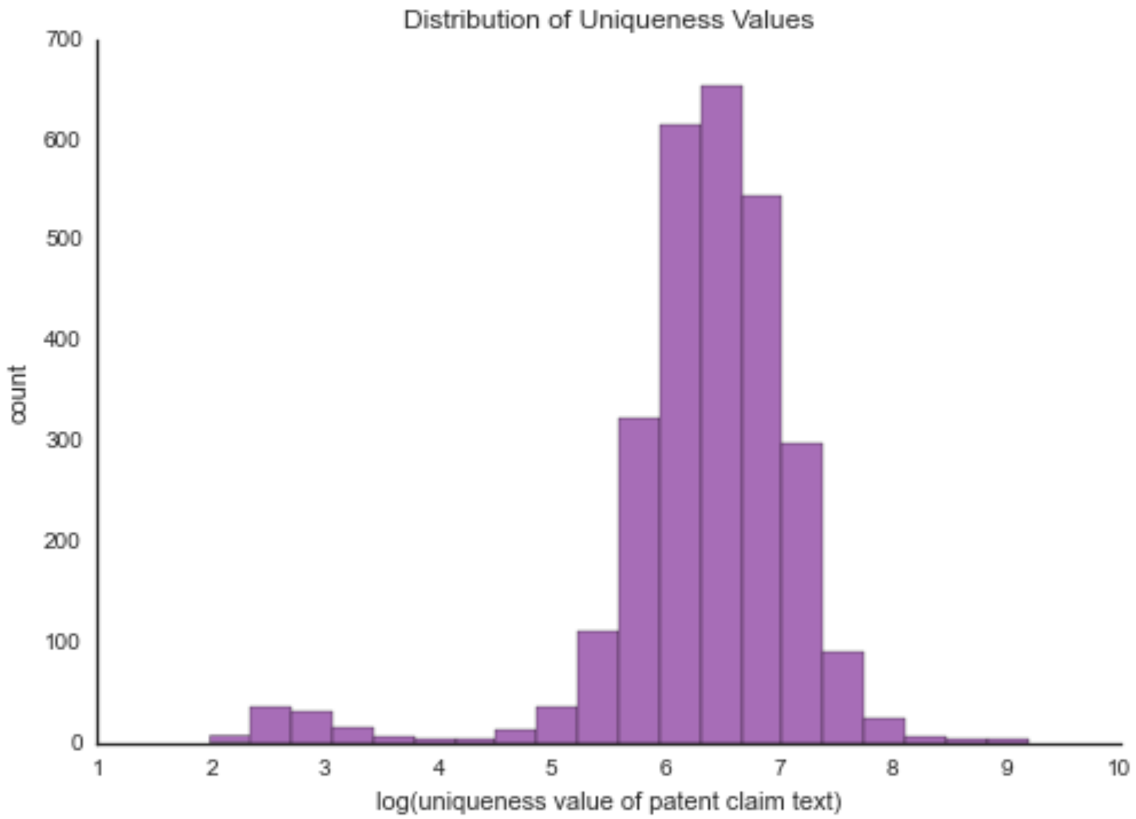
The final step was matching particular patents with particular funding rounds, using the company name as the key matching the two. As the focus of this research was startups, only companies with funding rounds labeled ‘angel,’ ‘seed,’ and ‘venture’ were selected. The earliest of these was then matched with the patent from the same company with the application date closest to that of the funding round. The patent and funding round were then merged to form one record.

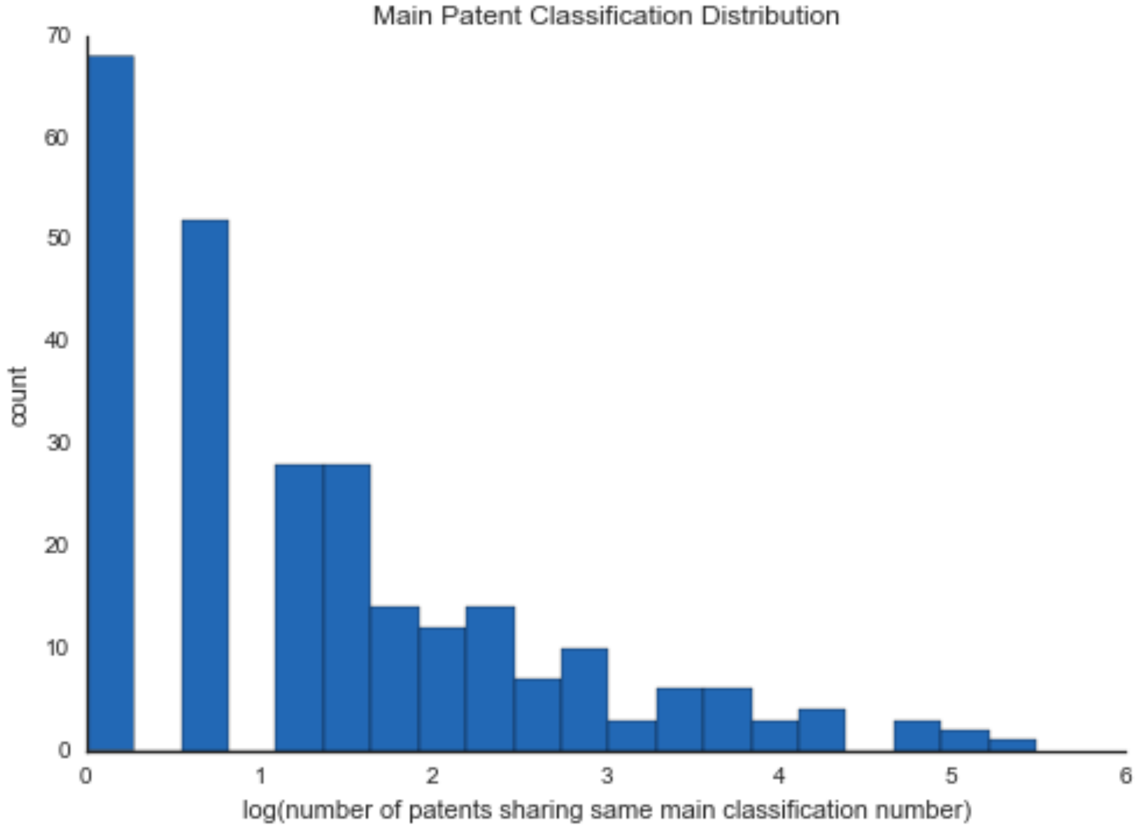
[5] Travis Lybbert and Nikolas Zolas, “Getting Patents and Economic Data to Speak to Each Other: An ‘algorithmic links with probabilities’ approach for joint analyses of patenting and economic activity,” Center for Economic Studies, vol. 12, iss. 16, September 2012; Ulrich Schmoch et al, “Linking Technology Areas to Industry Sectors,” Final Report to the European Commission, DG Research, November 2003.

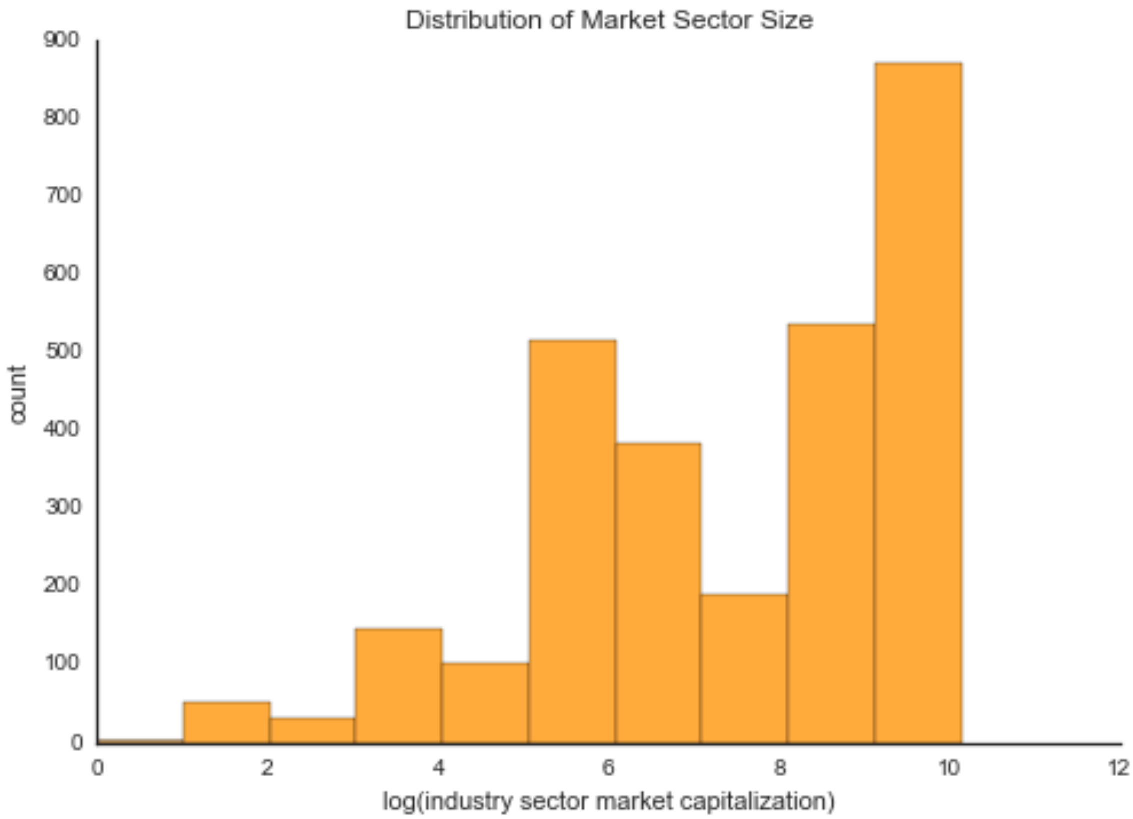
Data Exploration

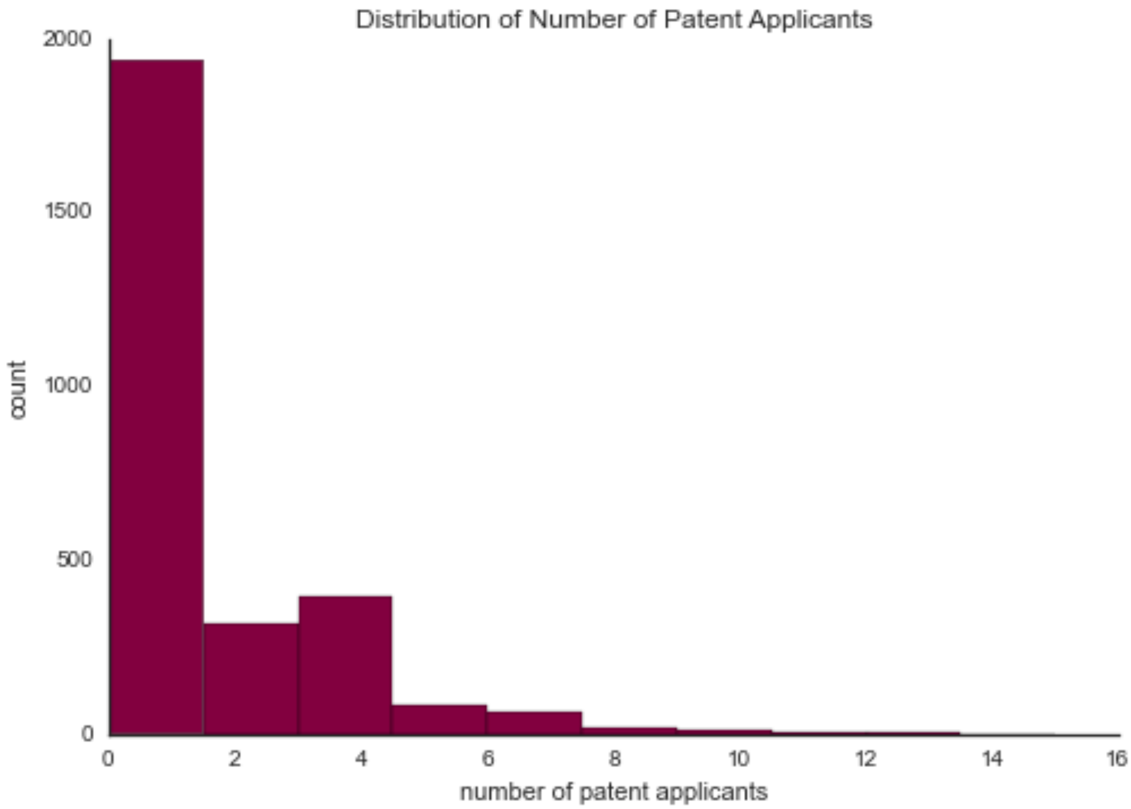
The initial, merged dataset of startup companies contained 2832 observations over 24 features ($n = 2832$, $p = 24$) and one response, the funding amount. Some variables were discarded because they were merely descriptive of other variables (e.g., the main classification title, the market sector name), dropping the number of features to seven ($p = 7$). Visualizations of the distributions of key variables follow.

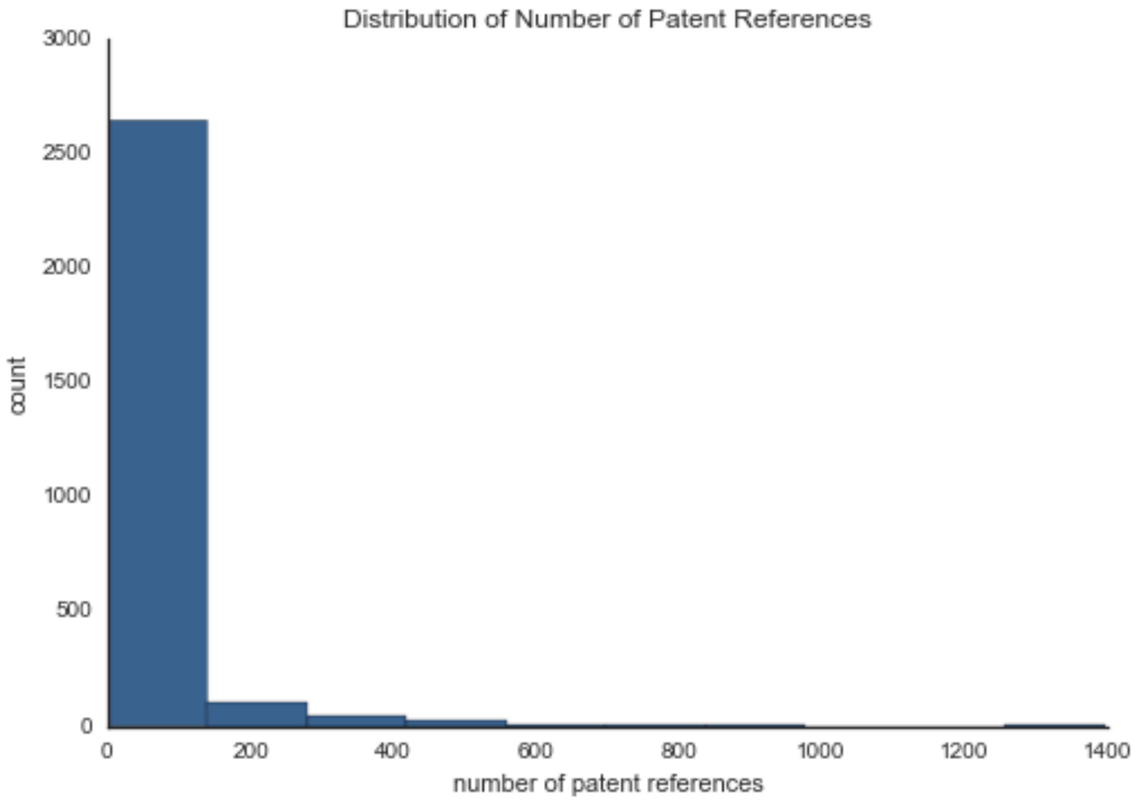



Distribution of Uniqueness Values
count
700
600
500
400
300
200
100
0
1
2
3
4
5
6
7
8
9
10
log(uniqueness value of patent claim text)


Main Patent Classification Distribution
count
70
60
50
40
30
20
10
0
0
1
2
3
4
5
6
log(number of patents sharing same main classification number)


Distribution of Market Sector Size
count
900
800
700
600
500
400
300
200
100
0
0
2
4
6
8
10
12
log(industry sector market capitalization)


Distribution of Number of Patent Applicants
count
2000
1500
1000
500
0
0
2
4
6
8
10
12
14
16
number of patent applicants



Analysis

Given the unclear relationship between the response (the amount of funding received) and any of the features, random forest regression was chosen as the primary modeling procedure. The random forest method is well suited to the task due to its capacity to handle nonparametric variables and its ability to identify the strengths of the different features.

The first approach utilized the log of the funding amount as the response; the features consisted of patent-specific variables, namely uniqueness value, market sector size, number of applicants, number of references, and the main classification of the patent. In addition, other categorical variables were included that pertained to the company, the state and country where the company is located. The company's location was included to see if that is related to the amount of funding received, e.g., if companies from California are better funded than those that are not. Dummy variables were generated to accommodate the categorical variables (n = 2832, p = 364). The optimal parameter *n_estimator* = 600 for the random forest regressor (*n_estimator* sets the number of trees utilized) was identified using ten-fold grid search cross-validation scored by mean squared error.

The results of the first approach were an out-of-bag (OOB) score of -0.1147764883222 and a root means square error (RMSE) of 1.71602477, slightly above one standard deviation of the response (1.668907). As is shown below, the uniqueness value was the only feature of any measurable importance:

First Startups Approach: n = 2832, p = 364	
Features	**Importance**
Uniqueness	0.274470
Number of References	0.067614
Market Sector Size	0.060545
Number of Applicants	0.037062
From California	0.018003
Patent Classification 424: Drug, bio-affecting and body treating compositions	0.011930
From New York	0.009787
From the US	0.009556
Patent Classification 707: Data Processing: database and file management or data structures	0.009417
From Texas	0.009222

Given the lackluster model performance, the next phase of analysis limited the features to the uniqueness value, market sector size, number of applicants, and number of references (*n* = 2832, *p* = 4). Ten-fold grid search cross-validation was again used to identify the optimal *n_estimator* parameter as well as the optimal maximum number of leaf nodes for each tree in the model. With *n_estimator* set to 800 and maximum leaf nodes set to six, the resulting OOB score was 0.002180985911123 and RMSE was 1.630339844022, slightly below one response standard deviation. This marked a marginal improvement over the first approach. The feature importance scores, shown below, generally followed what was expected given the research on what venture capitalists consider when making investment decisions, namely that uniqueness and market sector size are the main factors.

Second Startups Approach: n = 2832, p = 4	
Features	**Importance**
Uniqueness	0.374133
Market Sector Size	0.323855
Number of References	0.183901
Number of Applicants	0.118112

The thought occurred that perhaps the response variables were too dissimilar. Funding levels vary greatly from the very beginning angel/seed rounds, when money is raised to develop a product and conduct market research, to the first venture round, Series A, when money is raised to take the product or service to market. Observations that reported Series A funding rounds were separated into a new dataset with 925 observations, seven initial features, and one log-transformed response (*n* = 925, *p* = 7). Dummy variables were generated for the categorical features, resulting in *p* = 252.

The same process was repeated with a grid search cross-validation to determine the appropriate *n_estimators* parameter. With *n_estimators* set to 700, the OOB score for the more limited dataset was -0.1275947876465 and RMSE was 1.007061651451 (one standard deviation of the new response values was 0.978134). Uniqueness was again the only feature of any significant importance.

First Series A Approach: n = 925, p = 252	
Features	**Importance**
Uniqueness	0.269596
Market Sector Size	0.068458
Number of References	0.058733
Number of Applicants	0.040417
From California	0.020101
Patent Classification 290: Prime-mover dynamo plants	0.016295
Patent Classification 707: Data Processing: database and file management or data structures	0.015874
Patent Classification 435: Chemistry: molecular biology and microbiology	0.015526
From Colorado	0.014197
From India	0.012590

The final attempt mimicked the second process, limiting the features to uniqueness value, market sector size, number of applicants, and number of references (*n* = 925, *p* = 4), and grid search cross-validation of *n_estimators* and maximum leaf nodes, valued at 100 and two respectively. This resulted in an OOB score of -0.00289963267832 and RMSE of 0.962671348378, again a slight improvement over the larger feature set. Interestingly, the ranked importance of uniqueness and market sector size switched, with the importance of market sector size rising substantially.

Second Series A Approach: n = 925, p = 4	
Features	**Importance**
Market Sector Size	0.47
Uniqueness	0.39
Number of References	0.13
Number of Applicants	0.01

Challenges and Successes

The greatest challenges presented by this project were processing and cleaning the patent data. Much time was spent getting the xml's in working order, and still more was spent addressing data integrity. No satisfactory method of dealing with multiple spellings of the same text field was discovered, but an adequate method of handling the multiple main classification formats/misprints was determined.

The patent claim uniqueness value also displays promise. By expanding the corpus with more patents and developing a methodology for dealing with misspelled tokens in the text (an issue not attempted during this project), this variable may prove a good measure for determining patent novelty.

Possible Extensions and Applications

There are numerous ways that this research could be extended and improved, including:

- Applying the same methodology to patent application data rather than patent grants, as was done here. Given the sometimes decades-long lag between patent application and patent grant, patent applications would be a timelier data source.
- Comparing companies with patents that received funding to companies with patents that did not. This would provide a more rigorous estimation of patent value, although the data on unfunded companies may be difficult to acquire.
- Comparing the funding received by companies with patents to the funding levels of companies without patents. Not all startups operate in industries that lend themselves to patentable ideas or products (e.g., education), and it would be interesting to discover what features are relevant to funding levels in those cases.

There are also many ways this process could be applied to business. Startups could gauge how interesting a patent would be to venture capitalists. Venture capital firms could use it as a metric to vet investment opportunities and generate business intelligence about the funding activities of their competitors. Finally, policymakers and the local, regional, and national levels could use a model similar to this one to better understand developments in the economy and develop policies aimed at fostering innovation, either in general or in specific industry sectors.

Conclusion

Although the models developed over the course of this project did not perform well, they did demonstrate the potential of using the text of patent claims to develop a uniqueness value of a given patent. Random forest regression is also a worthwhile model for addressing research questions of this nature, and may end up performing relatively well with improved data cleaning and acquisition techniques.