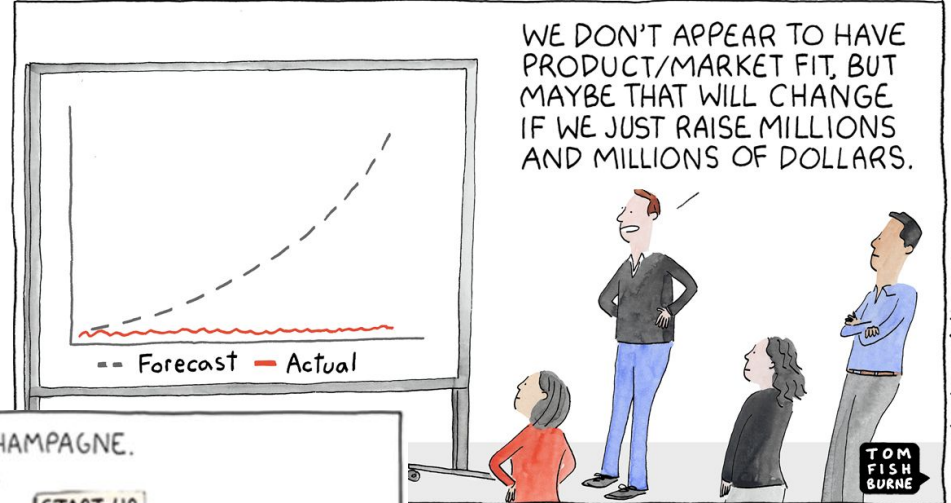


# Predicting Startup Success



## Startup



Uber



## Regular New Company



The main information we have on a new company are who founded it, what they're trying to do (company description, industry tags), location, name, logo, and website.

Can this tell me:

1. How **innovative** it is? → Assume so, compare it with other companies' descriptions
2. If the idea is **scalable**? → Nope
3. If it will ultimately be **successful**? → This is the goal, use the innovative-ness score along with other information

# Overview of the dataset without descriptions:

```
companies.head()
```

	Name	funding_total_usd	status	country_code	city	funding_rounds	founded_at	first_funding_at	last_funding_at	categories
0	#fame	10000000.0	operating	IND	Mumbai	1	2011-07-17 06:36:17.276987120	2015-05-01	2015-05-01	[Media]
1	:Qounter	700000.0	operating	USA	Delaware City	2	2014-04-09 00:00:00.000000000	2014-01-03	2014-10-14	[Application Platforms, Real Time, Social Netw...]
2	0-6.com	2000000.0	operating	CHN	Beijing	1	2007-01-01 00:00:00.000000000	2008-03-19	2008-03-19	[Curated Web]
3	01Games Technology	41250.0	operating	HKG	Hong Kong	1	2010-03-25 06:36:17.276987120	2014-01-07	2014-01-07	[Games]
4	Ondine Biomedical Inc.	762851.0	operating	CAN	Vancouver	2	1997-01-01 00:00:00.000000000	2009-11-09	2009-12-21	[Biotechnology]



- Created success column from combination of status and funding\_total\_usd columns
- Made binary column for country (USA vs. not)
- One-hot-encoded top cities and categories (14 & 50 respectively)
- Changed date columns to time since founding, time from founding to first funding, etc.
- Dropped name column
- Checked for multicollinearity and adjusted columns as necessary

# Overview of the dataset with descriptions:

desc.head()

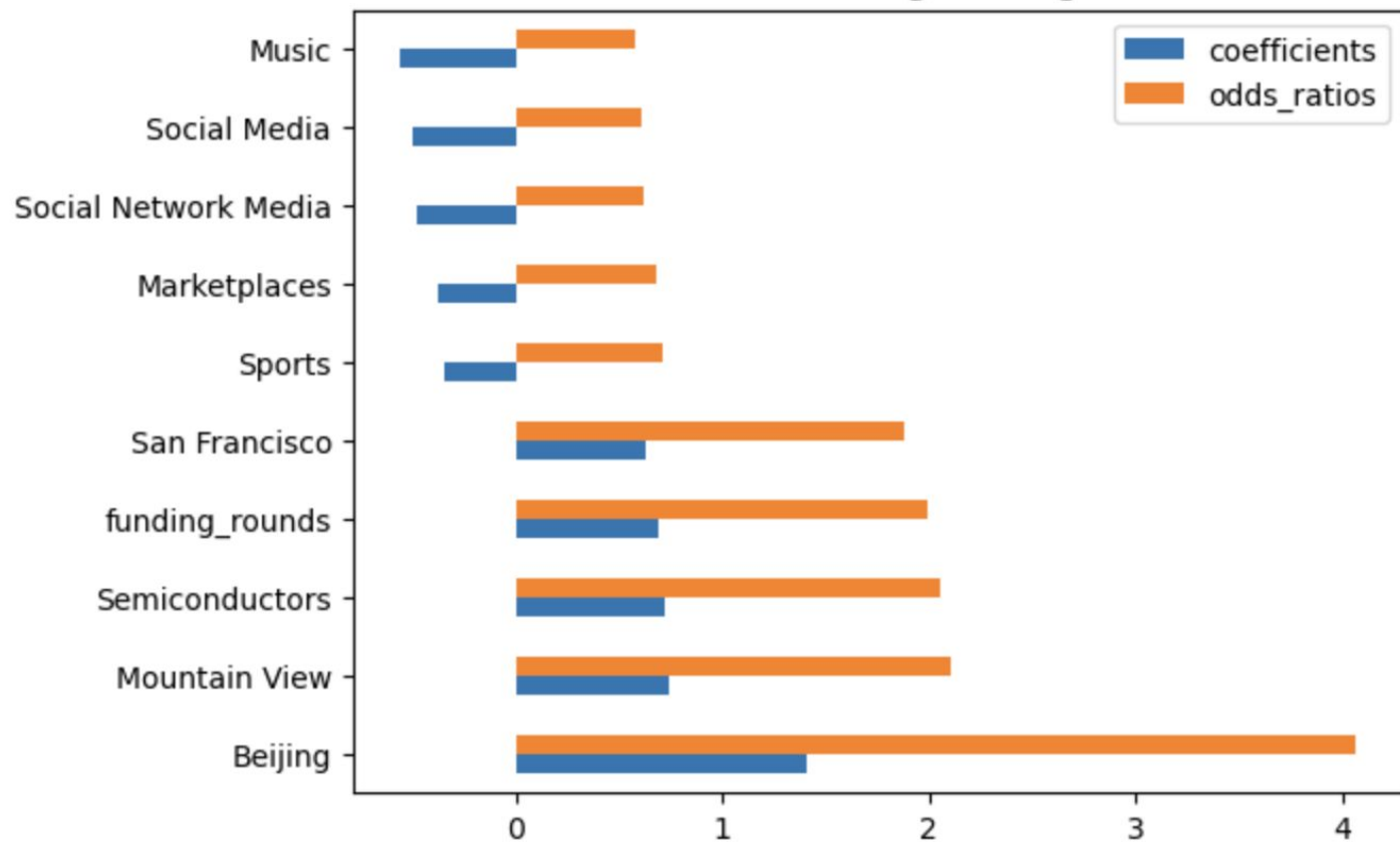
	name	city	tagline	description
0	Campus Bubble	Atlanta	Your Academic Identity	Campus Bubble ("CB") is the Academic Community...
1	DueProps	Atlanta	Gamifying the \$46 Billion Employee Incentives ...	t unprecedented ...
2	SalesLoft	Atlanta	Quickly build high-quality prospect lists	build high-quality prospect lists\n
3	The Coca-Cola Company	Atlanta	NaN	Coca-Cola Journey is a digital magazine that f...
4	EarthLink	Atlanta	NaN	NaN



- 1 Campus Bubble ("CB") is the Academic Community Network just as LinkedIn is the professional community network, and Facebook is the social community network. C  
B provides academic institutions with a student powered, cross platform, private online community, focused ...
- 2 t unprecedented ...
- 3 build high-quality prospect lists
- 4 t Coca-Cola Journey is a digital magazine that focuses on important topics, social causes and news about The Coca-Cola Company.
- 5 nan

	Manual train-test-split (train on companies older than 3.8 years, test on newer):	Auto train-test-split using scikit-learn function:
Logistic Regression	Training set accuracy: 72% Test set accuracy: 77% <b>Precision: 52%</b> Recall: 31% F1: 39%	Training set accuracy: 73% Test set accuracy: 73% <b>Precision: 75%</b> Recall: 66% F1: 70%
Decision Tree	Training set accuracy: 72% Test set accuracy: 78% Precision: 58% Recall: 23% F1: 33%	Training set accuracy: 73% Test set accuracy: 72% Precision: 74% Recall: 66% F1: 70%
Random Forest	Training set accuracy: 72% Test set accuracy: 74% Precision: 46% Recall: 50% F1: 48%	Training set accuracy: 74% Test set accuracy: 73% Precision: 73% Recall: 70% F1: 71%
Support Vector Machine	Training set accuracy: 71% Test set accuracy: 77% Precision: 54% Recall: 29% F1: 38%	Training set accuracy: 73% Test set accuracy: 73% Precision: 75% Recall: 65% F1: 70%
Neural Network	Highest accuracy: 78%	Highest accuracy: 73%

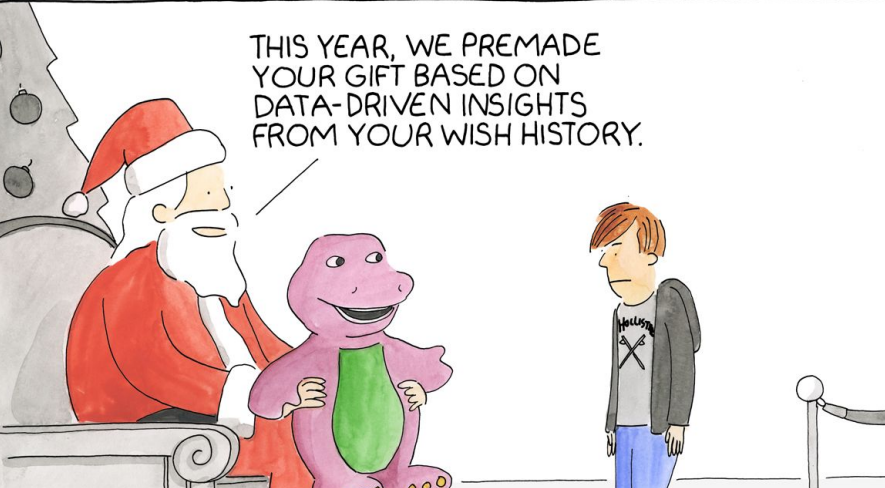
Coefficients from Logistic Regression



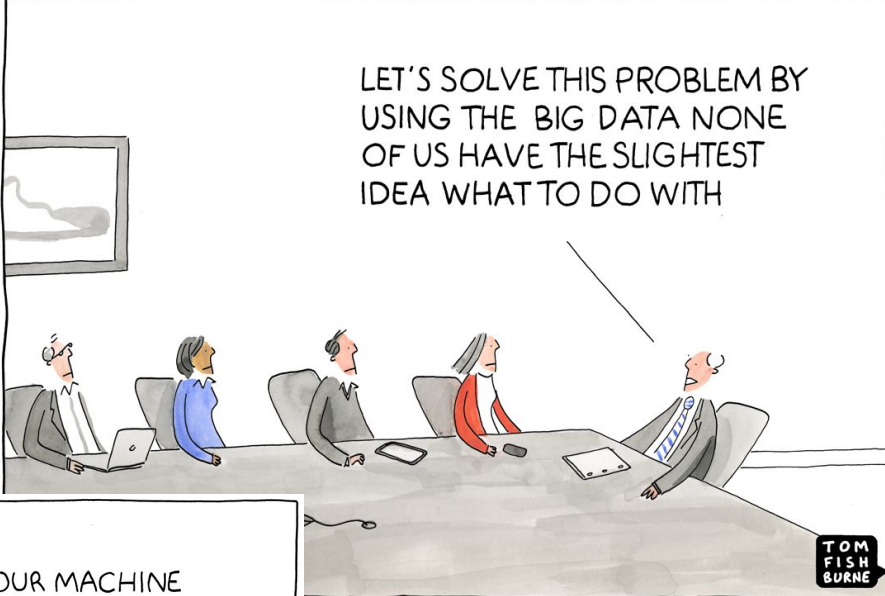
## Next Steps:

- Move on to Y Combinator data
- Get document word embeddings for descriptions
- Compute distance from most similar other company to get innovative-ness score
- Run models on this data





THIS YEAR, WE PREMADE  
YOUR GIFT BASED ON  
DATA-DRIVEN INSIGHTS  
FROM YOUR WISH HISTORY.



LET'S SOLVE THIS PROBLEM BY  
USING THE BIG DATA NONE  
OF US HAVE THE SLIGHTEST  
IDEA WHAT TO DO WITH

TOM  
FISH  
BURNÉ

©marketoost.com



SORRY, KID, OUR MACHINE  
LEARNING CRM WITH  
PREDICTIVE ANALYTICS SAYS  
YOU'RE GETTING COAL THIS YEAR.

TOM  
FISH  
BURNÉ

©marketoost.com