



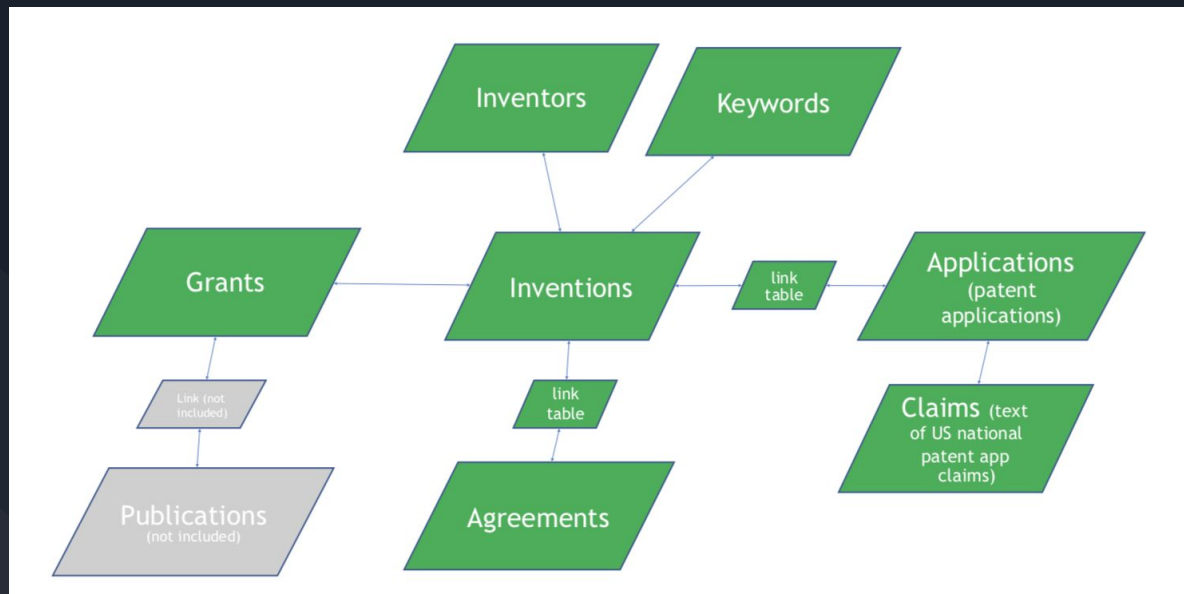
Tech-Transfer Analysis

Presented by

Mingfeng Li
Yinan Ling
Nan You
Jinhao Zhang

Columbia Technology Ventures

Entity-Relationship
Diagram:





Objective



Use the selected features and the model to classify the inventions into 6 categories by their abilities to make money

Which 6 categories of inventions

Divided by the revenue of their agreements with companies:

1. "A" - \$0
2. "B" - 0-\$10K
3. "C" - \$10K-\$100K
4. "D" - \$100K-\$1M
5. "E" - >\$1M
6. "Z" - did not obtain any agreement



Initial Analysis

The relevant features :



- Invention Key (serve as index)
- PeopleKey- The number of People Collaborating
- Department- The number of Departments Collaborating

Initial Analysis

The relevant features :



- fundingAmount - Total Funding amount received
- Type - Agreements Type
- countryCode - count of countries inventions applied for patents
- Application Status

Initial Analysis

The relevant features :



- Status - Application Status
 - a. abandoned ->0
 - expired ->0.5
 - filed ->0.5
 - issued ->1

Initial Analysis

- Data Cleaning

1. Join all the relevant features to one table



	inventionsKey	peopleKey	department	fundingAmount	type	revenue	countryCode	status
0	K2AD0872121812714669	4.0	2.0	3525947.0	MTA	A \$0	3	0.5
1	K2AD0872121812714669	4.0	2.0	3525947.0	other	A \$0	3	0.5
3	K2AD0872121812714669	4.0	2.0	3525947.0	CDA	A \$0	3	0.5
5	K2AD0872124416279053	1.0	1.0	NaN	NaN	NaN	0	0.0
6	K2AD0872124530352720	2.0	2.0	0.0	LicenseNonEx	C 10K–100K	5	2.5



Initial Analysis

- Data Cleaning

2. Change the data types to the ones that can be easily analyzed

	inventionsKey	peopleKey	department	fundingAmount	type	revenue	countryCode	status
0	K2AD0872121812714669	4.0	2.0	3525947.0	5	0	24	1
1	K2AD0872121812714669	4.0	2.0	3525947.0	11	0	24	1
3	K2AD0872121812714669	4.0	2.0	3525947.0	0	0	24	1
5	K2AD0872124416279053	1.0	1.0	NaN	10	5	0	0
6	K2AD0872124530352720	2.0	2.0	0.0	4	2	37	19

Initial Analysis

- Data Cleaning

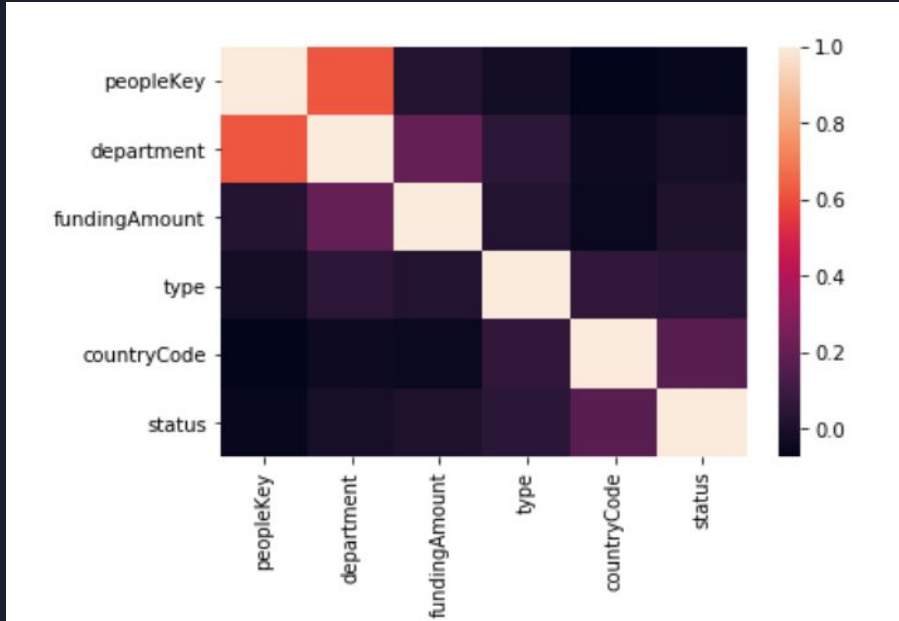
3. Normalize the data and take care of all the invalid data in the set

	inventionsKey	peopleKey	department	fundingAmount	type	revenue	countryCode	status
0	K2AD0872121812714669	0.25	0.272727	3.201032e-03	0.454545	0	0.571429	0.025641
1	K2AD0872121812714669	0.25	0.272727	3.201032e-03	1.000000	0	0.571429	0.025641
3	K2AD0872121812714669	0.25	0.272727	3.201032e-03	0.000000	0	0.571429	0.025641
5	K2AD0872124416279053	0.10	0.181818	0.000000e+00	0.909091	5	0.000000	0.000000
6	K2AD0872124530352720	0.15	0.272727	9.078501e-10	0.363636	2	0.880952	0.487179
7	K2AD0872124530352720	0.15	0.272727	9.078501e-10	0.363636	2	0.880952	0.487179

n=11,465

Initial Analysis

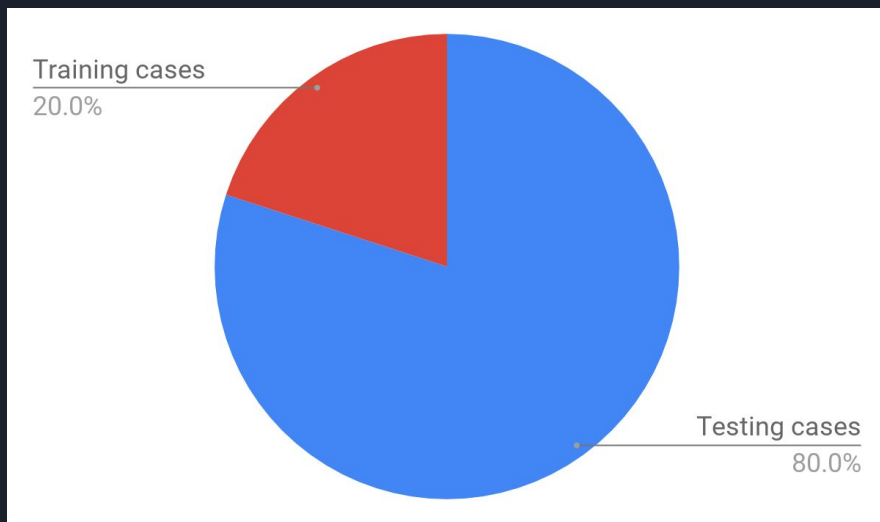
- Correlation Analysis



- Fairly strong correlation exists between the number of departments and the number of people collaborating, as expected

Machine Learning

- Splitting the data



- 80% of data is randomly chosen to be the training cases
- The remaining 20% is chosen to be the testing cases



Machine Learning

- Choice of Algorithm



Neural Network



Result

- Model Accuracy

92% Accuracy!!! in the first place

But.....

We realized that the set is biased because there are too many zeros in the “revenue” feature

Too many 0s means too many data falls in Category ‘A’



Result

- Model Improvement

- Apply balancing
- Balanced the batch data by selecting the same amount of data from each revenue category

Result

- Model Accuracy after Balancing
the data

78.7% accuracy
in test cases





Conclusion

- Use the selected features and the model to classify the inventions into 6 categories by their abilities to make money
- We trained a Neural Network model that has 78.7% of accuracy in prediction

Future Work



- A further in-depth analysis of each factor's commercial impact
- Take the time into consideration