Predicting Startups Success

Raghad Alarifi

Agenda

- Project Goal
- Data Set
- Classification Model Comparison
- Results

Project Goal

- Build a classification model to **predict startup business success**

Data Set

Crunchbase data set

Crunchbase is **a platform for finding business information about private and public companies**.

- Obtain Companies and Rounds 2013 Datasets from Crunchbase.com
- Companies Dataset : 17 feature and 17727 rows
- Rounds Dataset : 13 feature and 31679 rows

Companies

name	category_code	funding_total_usd	status	country_code	state_code	region	city	funding_rounds	founded_at	founded_month	founded_quarter	founded_year	first_funding_at	last_funding_at	last_m
Mocana	security	35500000	operating	USA	CA	SF Bay	San Francisco	4	2002-01-01	2002-01	2002-Q1	2002.0	2006-04-01	2012-08-23	
BombBomb	games_video	500000	operating	USA	CO	Colorado Springs	Colorado Springs	1	NaN	NaN	NaN	NaN	2013-05-30	2013-05-30	
Zoom Media & Marketing - United States	advertising	30000000	operating	USA	NY	New York	New York	1	1991-01-01	1991-01	1991-Q1	1991.0	2009-03-01	2009-03-01	
3D Sports Technology	software	404940	operating	USA	MN	Minneapolis	Minneapolis	2	2010-01-01	2010-01	2010-Q1	2010.0	2012-06-07	2013-07-25	
Spring Bank Pharmaceuticals	biotech	8100000	operating	USA	MA	Boston	Milford	2	NaN	NaN	NaN	NaN	2011-05-05	2013-01-25	

Rounds

company_name	company_category_code	company_country_code	company_state_code	company_region	company_city	funding_round_type	funded_at	funded_month	funded_quarter	funded_year	raised_amount_usd
PlantSense	web	USA	CA	SF Bay	San Francisco	venture	2010-07-19	2010-07	2010-Q3	2010	2500000.0
NanoSteel	nanotech	USA	RI	Providence	Providence	venture	2011-10-18	2011-10	2011-Q4	2011	17000000.0
Appistry	analytics	USA	MO	Saint Louis	St. Louis	series-c+	2009-01-01	2009-01	2009-Q1	2009	3000000.0
Gr8erMinds	software	USA	IN	Indianapolis	Indianapolis	angel	2011-07-06	2011-07	2011-Q3	2011	4000.0
CE Interactive	software	USA	NY	New York	New York	angel	2005-01-01	2005-01	2005-Q1	2005	NaN

Data Assumption

Target Variable : success

name	category_code	funding_total_usd	status	country_code	state_code	region	city	funding_rounds	founded_at	founded_month	founded_quarter	founded_year	first_funding_at	last_funding_at	last_m
Mocana	security	35500000	operating	USA	CA	SF Bay	San Francisco	4	2002-01-01	2002-01	2002-Q1	2002.0	2006-04-01	2012-08-23	
BombBomb	games_video	500000	operating	USA	CO	Colorado Springs	Colorado Springs	1	NaN	NaN	NaN	NaN	2013-05-30	2013-05-30	
Zoom Media & Marketing - United States	advertising	30000000	operating	USA	NY	New York	New York	1	1991-01-01	1991-01	1991-Q1	1991.0	2009-03-01	2009-03-01	
3D Sports Technology	software	404940	operating	USA	MN	Minneapolis	Minneapolis	2	2010-01-01	2010-01	2010-Q1	2010.0	2012-06-07	2013-07-25	
Spring Bank Pharmaceuticals	biotech	8100000	operating	USA	MA	Boston	Milford	2	NaN	NaN	NaN	NaN	2011-05-05	2013-01-25	

IPO or Acquired = **Succeed**

Closed or No funding in 2 years = **Failed**

Feature Engineering

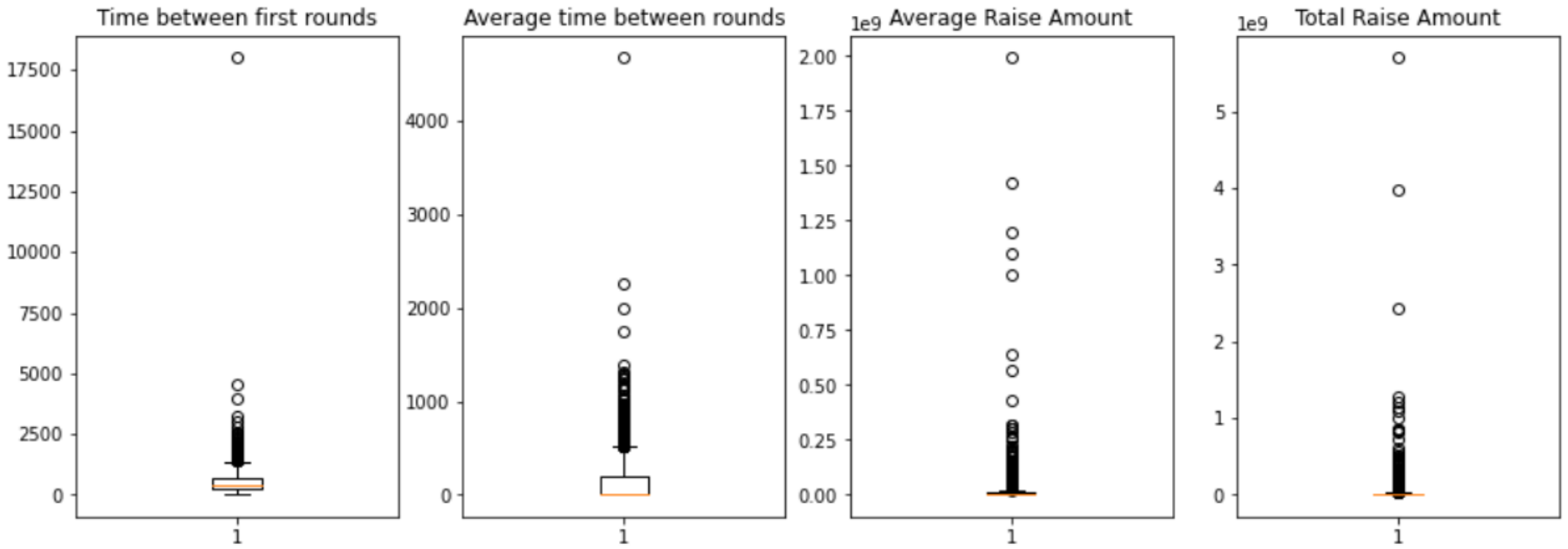
Process of using domain knowledge to extract features

- Generate class label
- Merge data sets to calculate : Average Time between Rounds , time between first and second round, avg_raise_usd
- Dummy Variables for categorial features

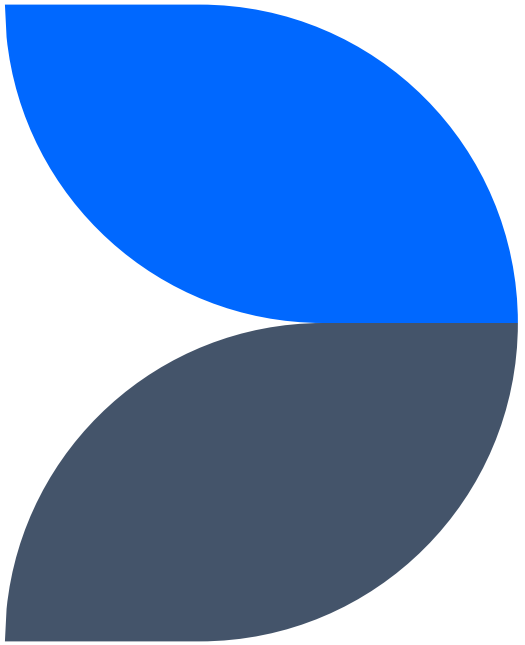
Data Cleaning

Drop	Drop	Replace
Drop Unnecessary Columns	Drop rows with Noisy Data	Replace outliers with column’s mean

Outliers


Time between first rounds
17500
15000
12500
10000
7500
5000
2500
0
1
Average time between rounds
4000
3000
2000
1000
0
1
1e9 Average Raise Amount
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
1
1e9 Total Raise Amount
5
4
3
2
1
0
1

Modeling

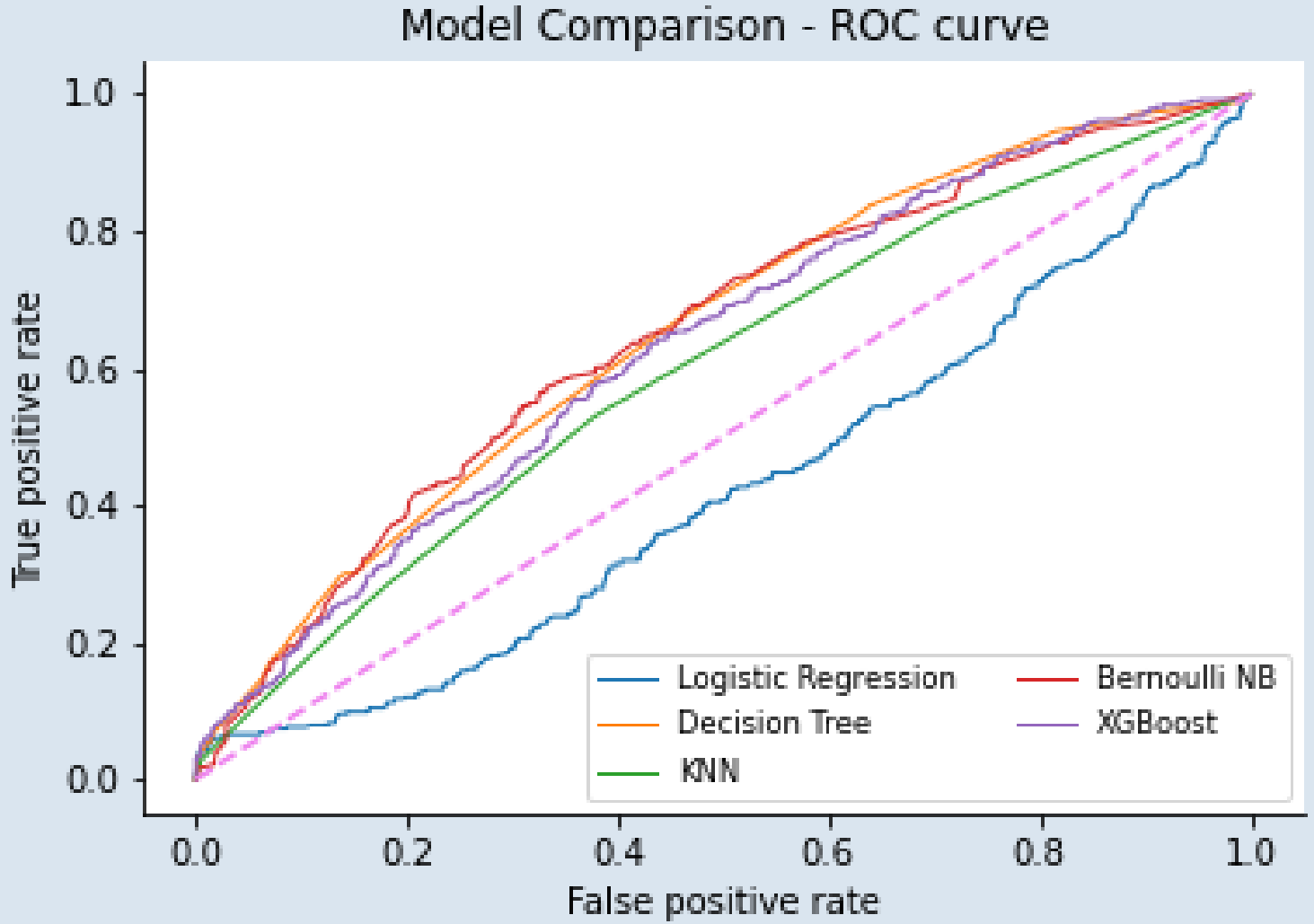


Class Distribution

Succeed = 30%

Failed = 70%

Comparing Models



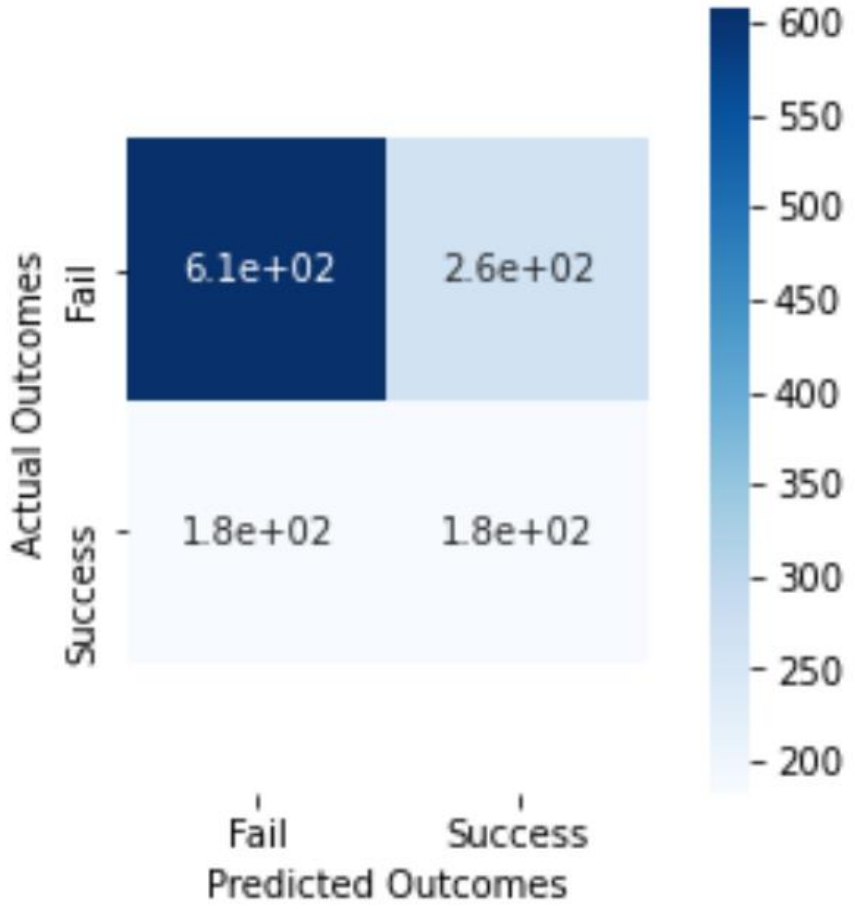
Model evaluation scores

```
ROC AUC score = 0.424223 for Logistic Regression
ROC AUC score = 0.650850 for Decision Tree
ROC AUC score = 0.597079 for KNN
ROC AUC score = 0.648985 for Bernoulli Naive Bayes
ROC AUC score = 0.633905 for XGBoost
```

Model evaluation scores

```
f_beta score = 0.807590 for Logistic Regression
f_beta score = 0.802441 for Decision Tree
f_beta score = 0.714625 for KNN
f_beta score = 0.807590 for Bernoulli Naive Bayes
f_beta score = 0.808262 for XGBoost
```

Confusion Matrix



Thank you