



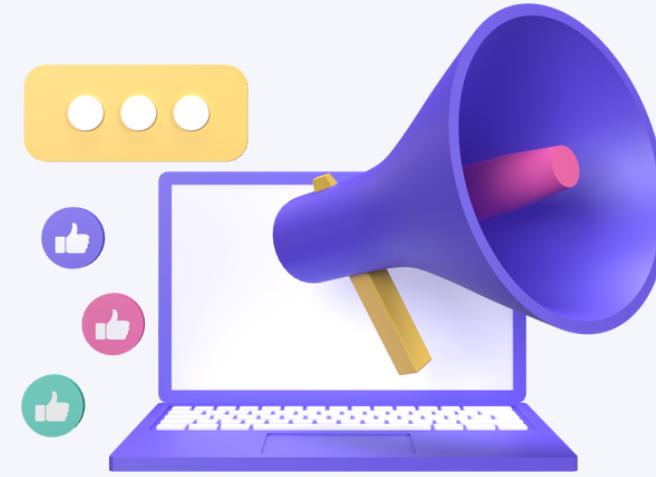
PREDICTING START-UP SUCCESS

By: Apurva Audi, Amanda Nguyen, James Anderson,
Sameer Ahmed, and Shubhada Kapre

PROBLEM STATEMENT

We ultimately attempt to predict whether or not a company in a top US start up city will be successful.

AGENDA



Introduction

Background about the dataset



Analysis

How we approached this problem



Framework

What methods we used to analyze data



Results

What we found from our analysis

INTRODUCTION



What we are predicting:

Prediction of start-up success based on:

- Total funding
- Number of funding rounds
- Size of funding : seed & venture
- Market: Technology & Non-technology
- Time between first and last funding round

Inherent Problems:

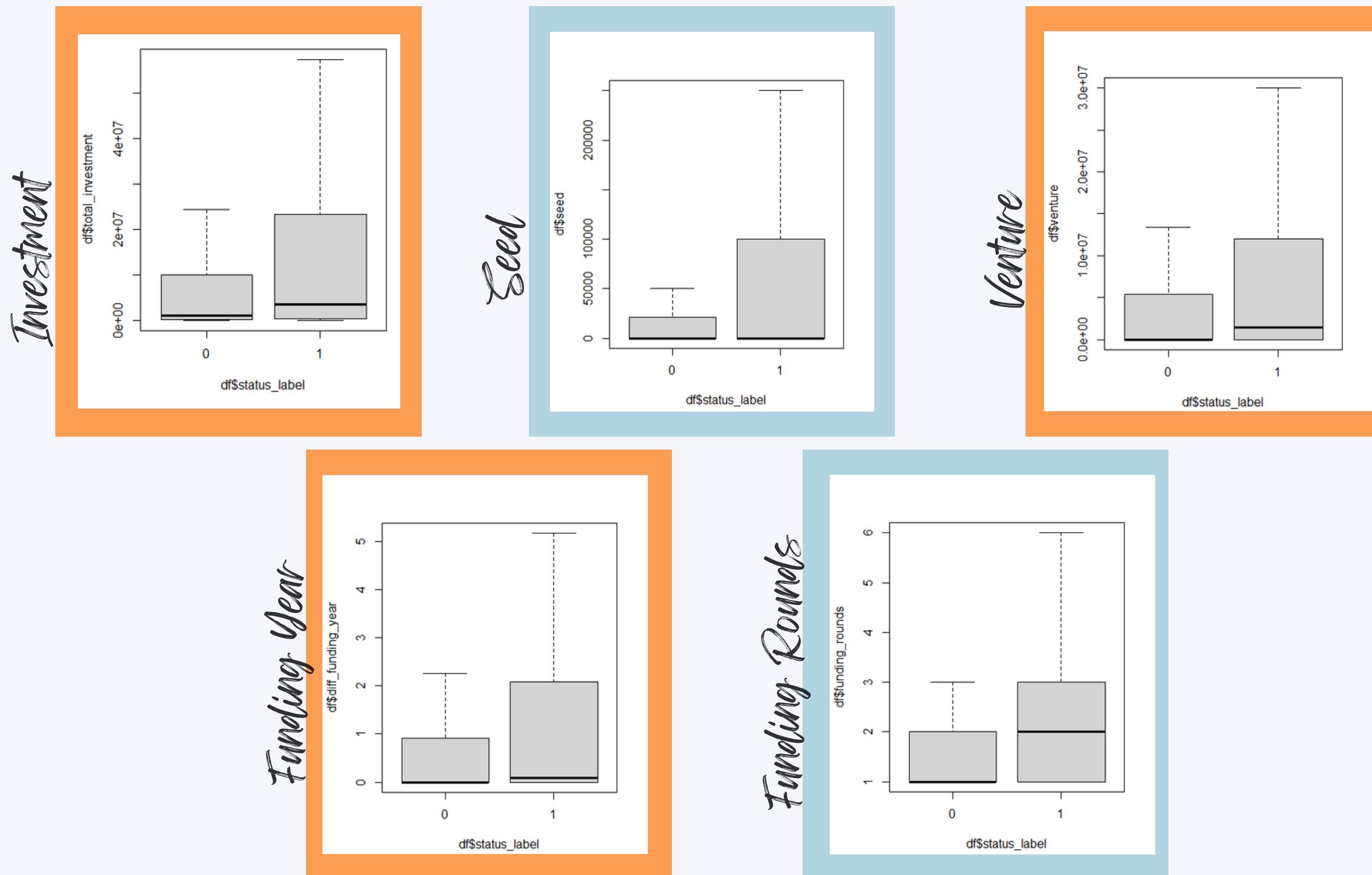
- Defining success vs failure is incredibly nuanced.
- Data given classified companies as "operating", "acquired", or "closed".
- 5% of data set was "closed", a strong bias given that in reality, 90% of start ups fail

More PROBLEMS

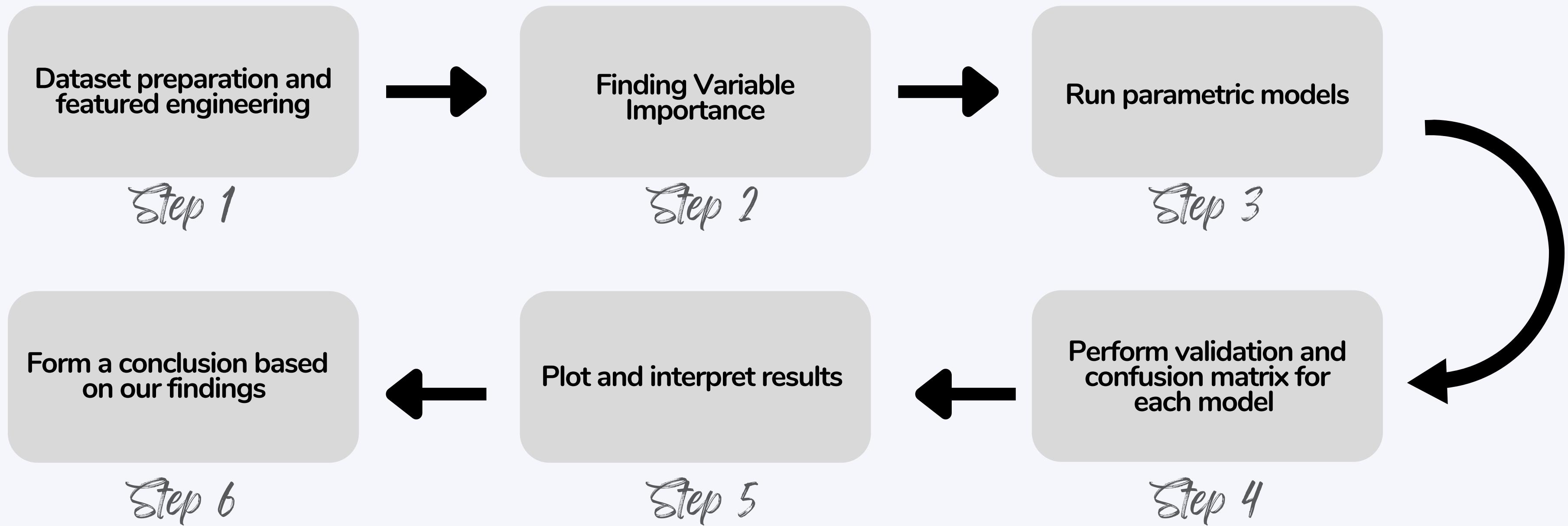


- Unrealistic closed company data set
- Operating & Acquired are subjective indicators of success
- Different companies have different funding needs
- Models do not predict seed-stage companies
- Region/Industry heavily influence financial data

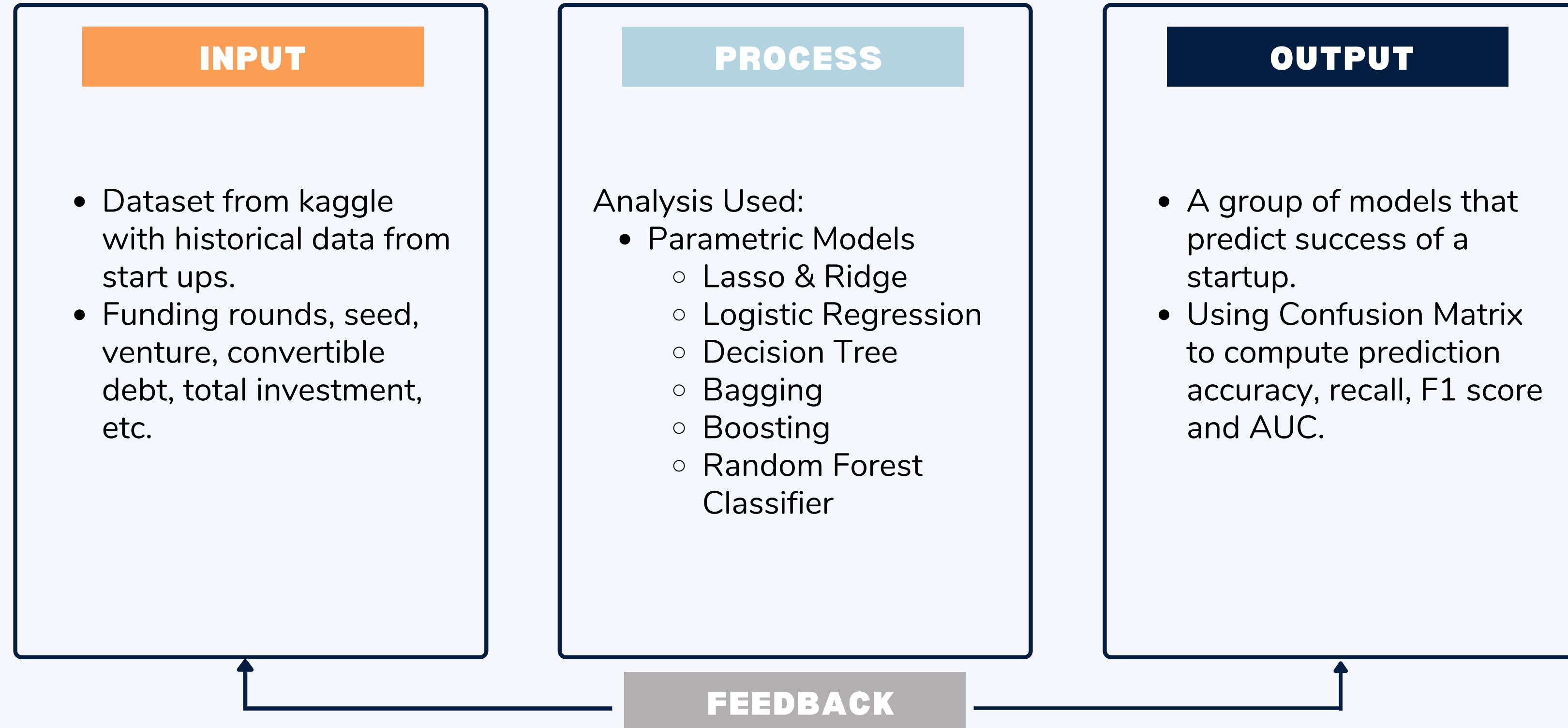
EXPLORATORY DATA ANALYSIS



FRAMEWORK

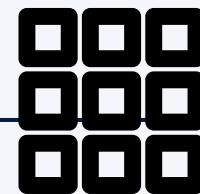
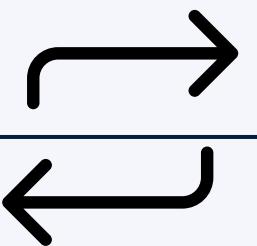
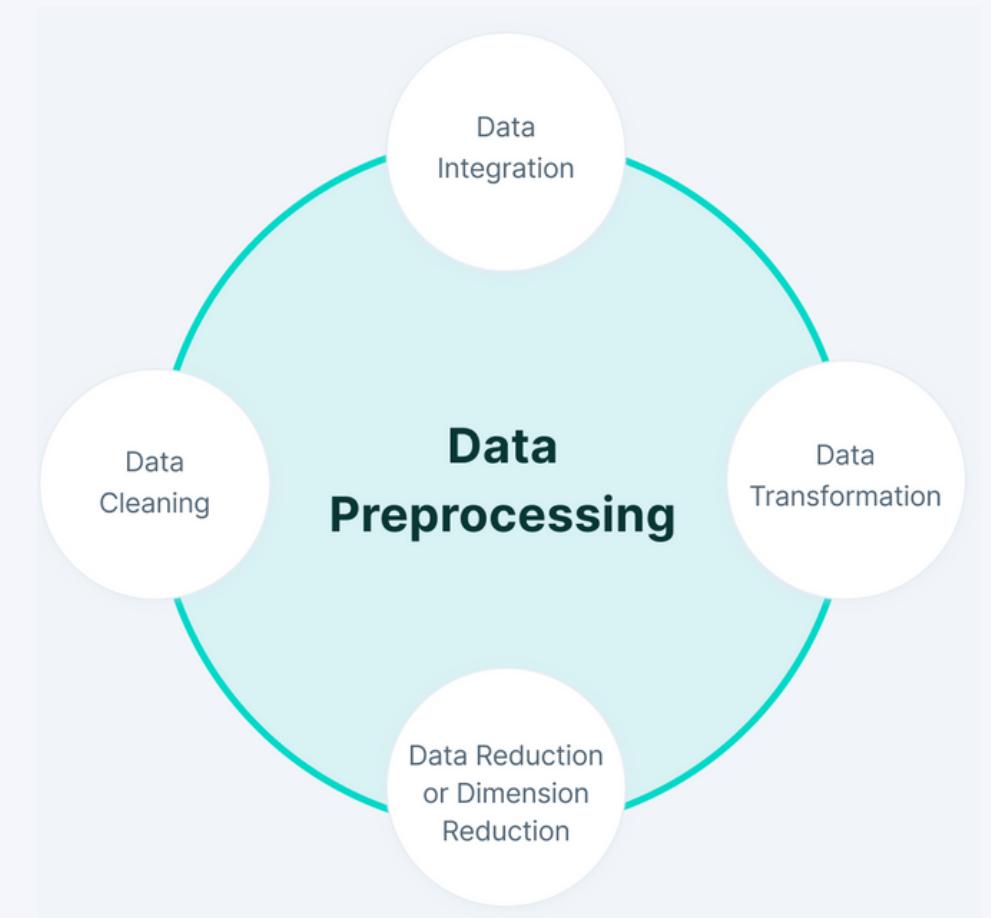


FRAMEWORK



DATA PRE-PROCESSING

- Actual records in the dataset - **44k**
- Covers startups across **120+** countries and **700+** markets
- Focused analysis on USA market in top **8** startup areas



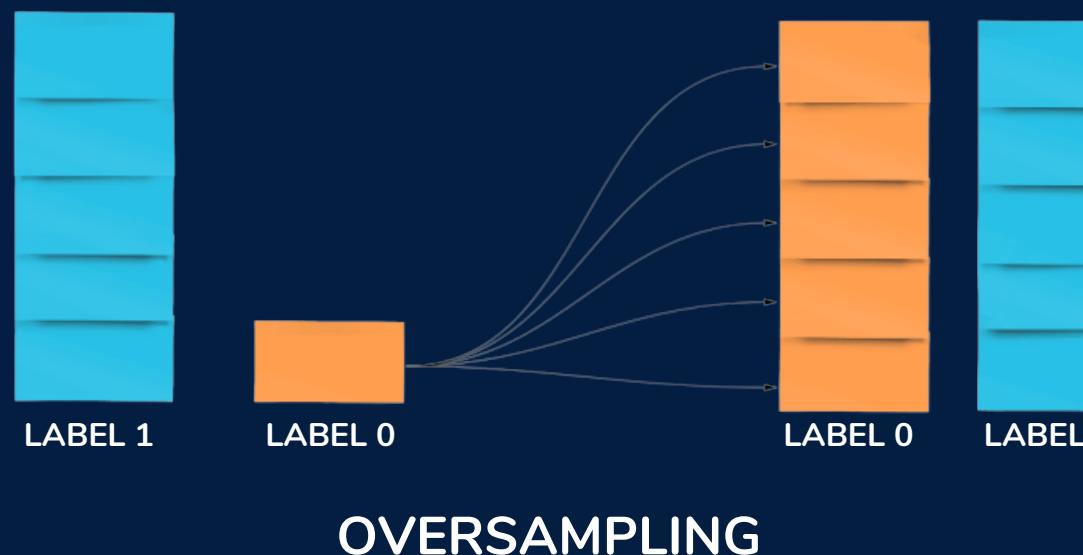
**Removing records
with missing values.**

**Converting strings
to numerical values
for funding**

**Classifying Market :
Technology & Non-
Technology**

**Handling
Categorical Data by
creating Categorical
Bins and using
numerical encoding**

IMBALANCE CLASSES IN DATASET



Dataset before Over Sampling

```
> table(df_inv$status_label)
```

0	1
716	12095

Using Over Sampling via the package **ROSE** - Random Over-Sampling Method to replicate the minority class

Dataset after Over Sampling

```
> data_balanced_over <- ovun.sample(status_label ~ ., data = df_inv, method = "over", N = 24190)$data
```

```
> table(data_balanced_over$status_label)
```

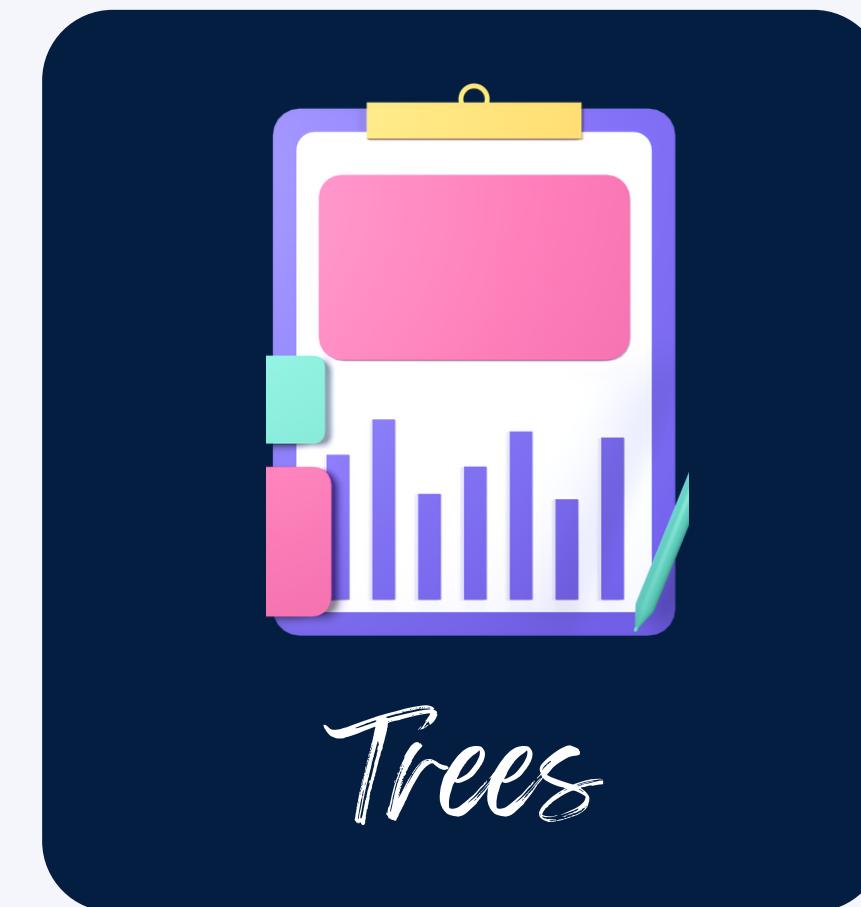
0	1
12095	12095

*Status Label:

0 - Failure (close)

1 - Success (acquired, operating)

ANALYSIS MODELS

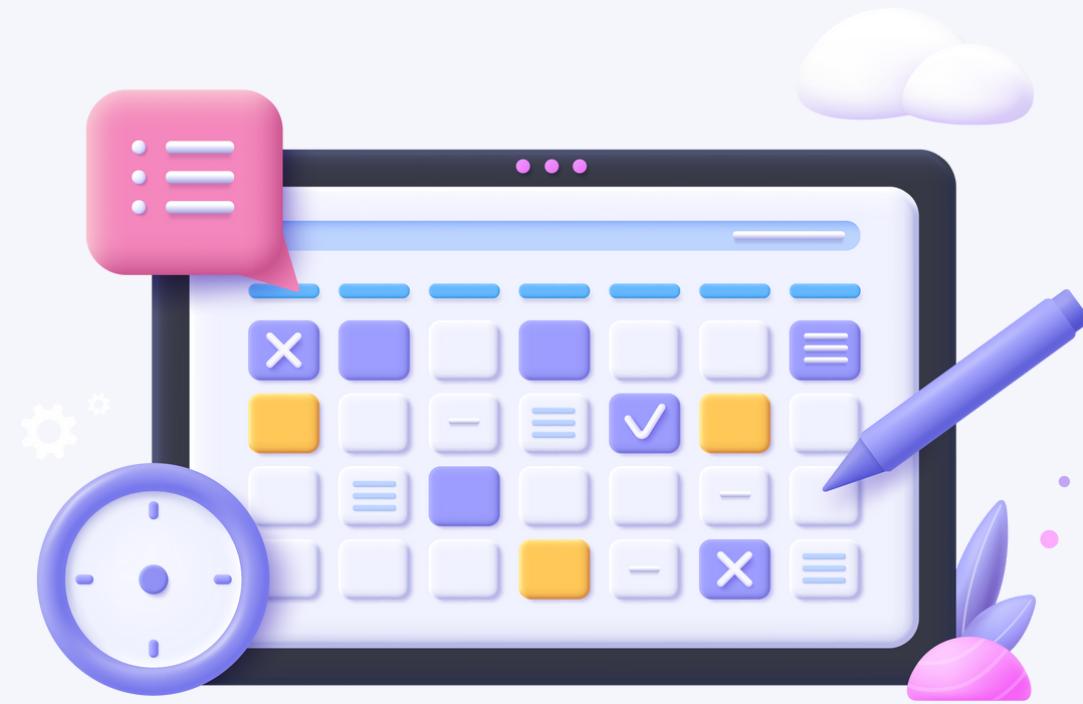


VARIABLE SELECTION



Lasso Regression

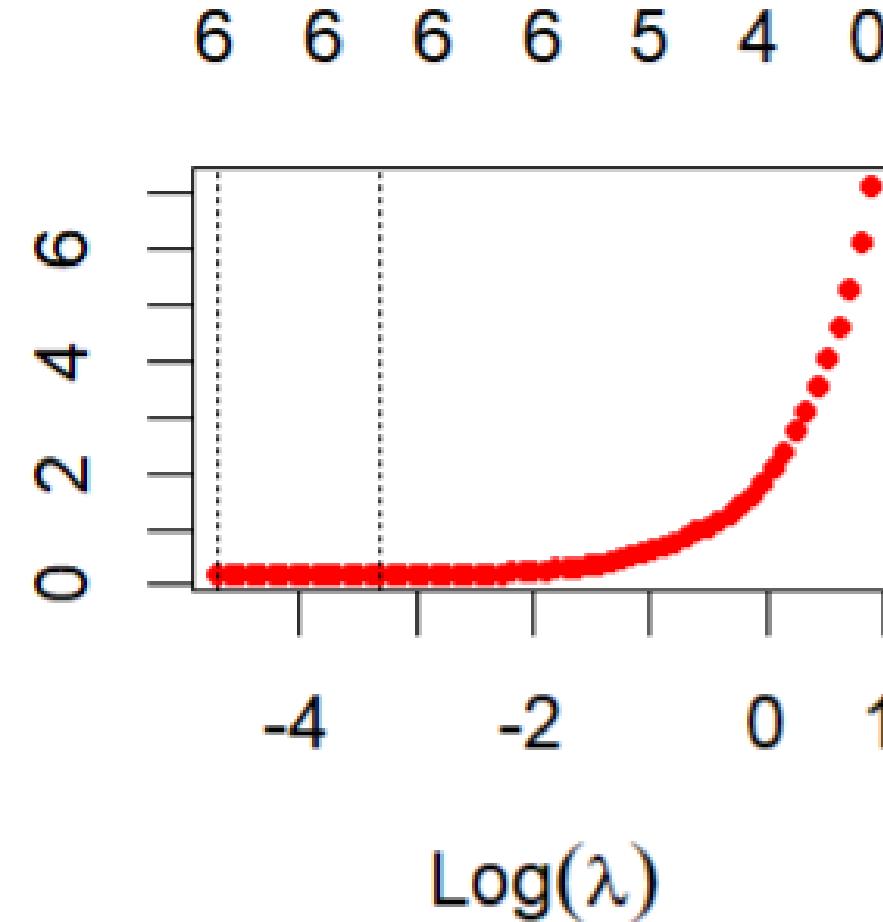
For removing non-important predictors
from the model



Ridge Regression

Assuring penalty factor and coefficients
are appropriate for accurate prediction

Mean-Squared Error

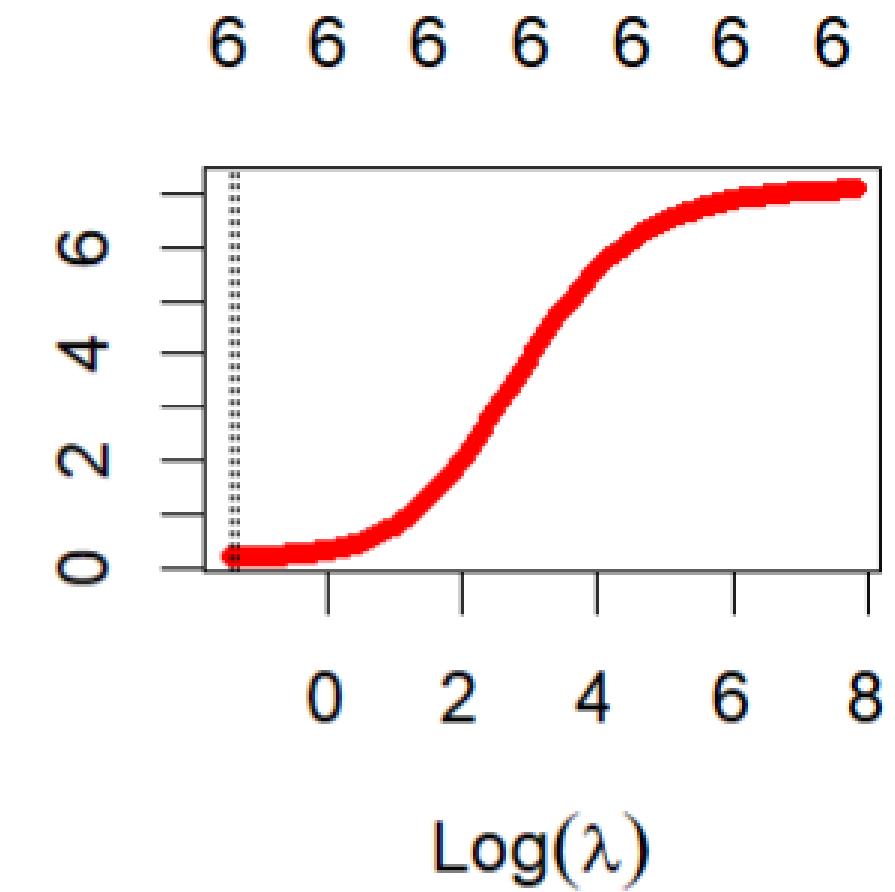


$\text{Log}(\lambda)$

LAMBDA SELECTION

determining the right amount of shrinkage

Mean-Squared Error



$\text{Log}(\lambda)$

Lasso Regression

(Intercept)		market_type	
	1.0117284		0.9755215
total_investment_cde	1.0550909	diff_funding_year_cde	1.0088068
funding_rounds_cde	0.9741143	seed_cde	0.9537323
venture_cde	1.0134712		

Ridge Regression

(Intercept)		market_type	
	1.1613541		0.9312577
total_investment_cde	0.9604958	diff_funding_year_cde	1.0314023
funding_rounds_cde	0.9928223	seed_cde	0.9005296
venture_cde	1.0350786		

VARIABLE COEFFICIENTS



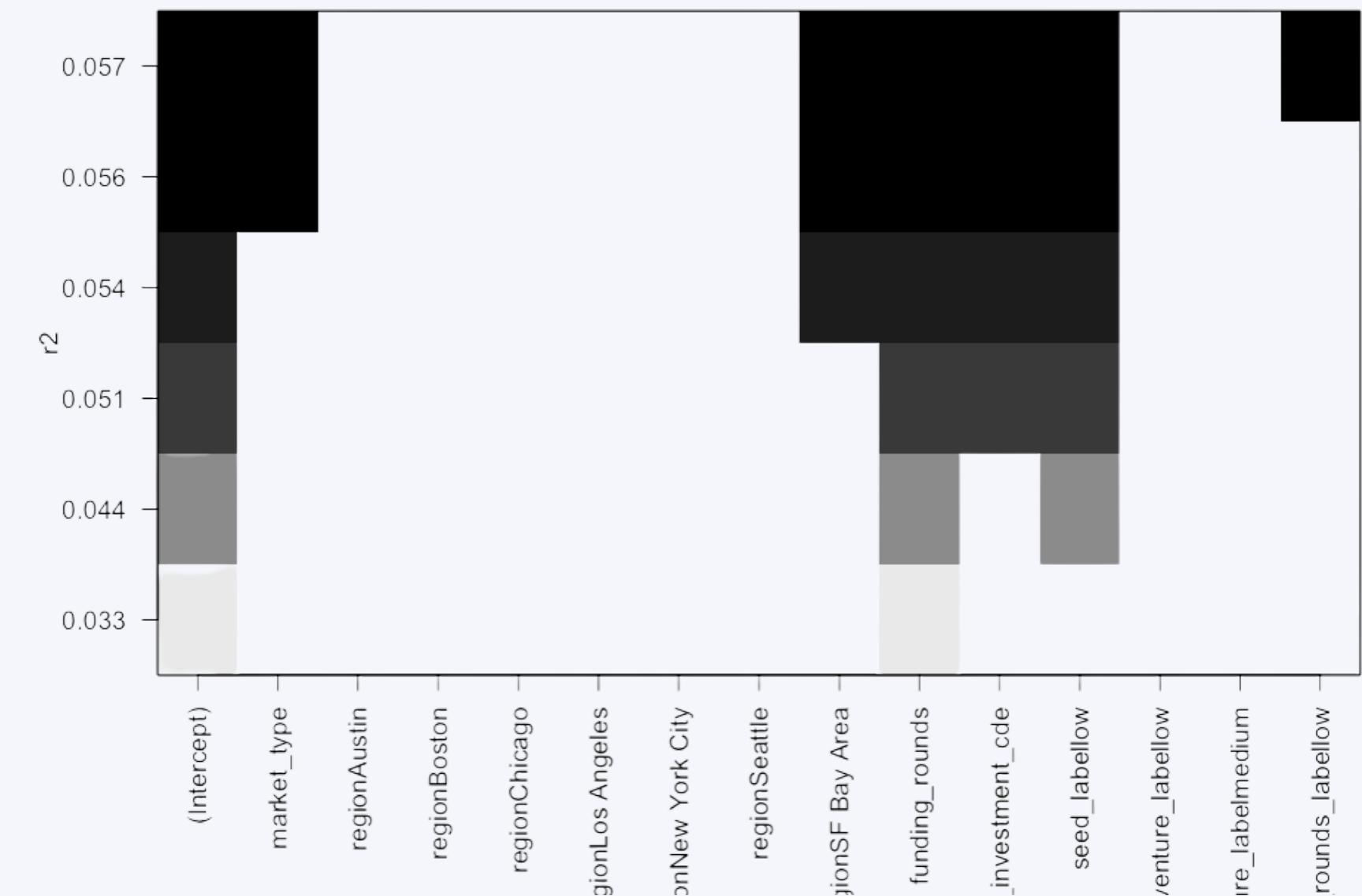
Logistic Regression

FEATURE SELECTION

2 MAIN FACTORS:

Lasso and Ridge
feature selection

Regsubset backward
selection



OUR BEST LOG MODEL

TOTAL INVESTMENT

FUNDING ROUNDS

SEED

VENTURE

MARKET TYPE

FUNDING YEAR

```
Call:  
glm(formula = status ~ diff_funding_year + market_type + funding_rounds +  
    seed_label + total_investment_cde, family = "binomial", data = df_inv_Train)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.0029	-1.0775	0.5758	1.1569	1.4283

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.12912	0.05490	2.352	0.0187 *
diff_funding_year	0.08910	0.01654	5.386	7.19e-08 ***
market_type	-0.22241	0.03301	-6.738	1.61e-11 ***
funding_rounds	0.11410	0.02025	5.636	1.74e-08 ***
seed_labellow	-0.59355	0.04389	-13.522	< 2e-16 ***
total_investment_cde	0.16658	0.01756	9.488	< 2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

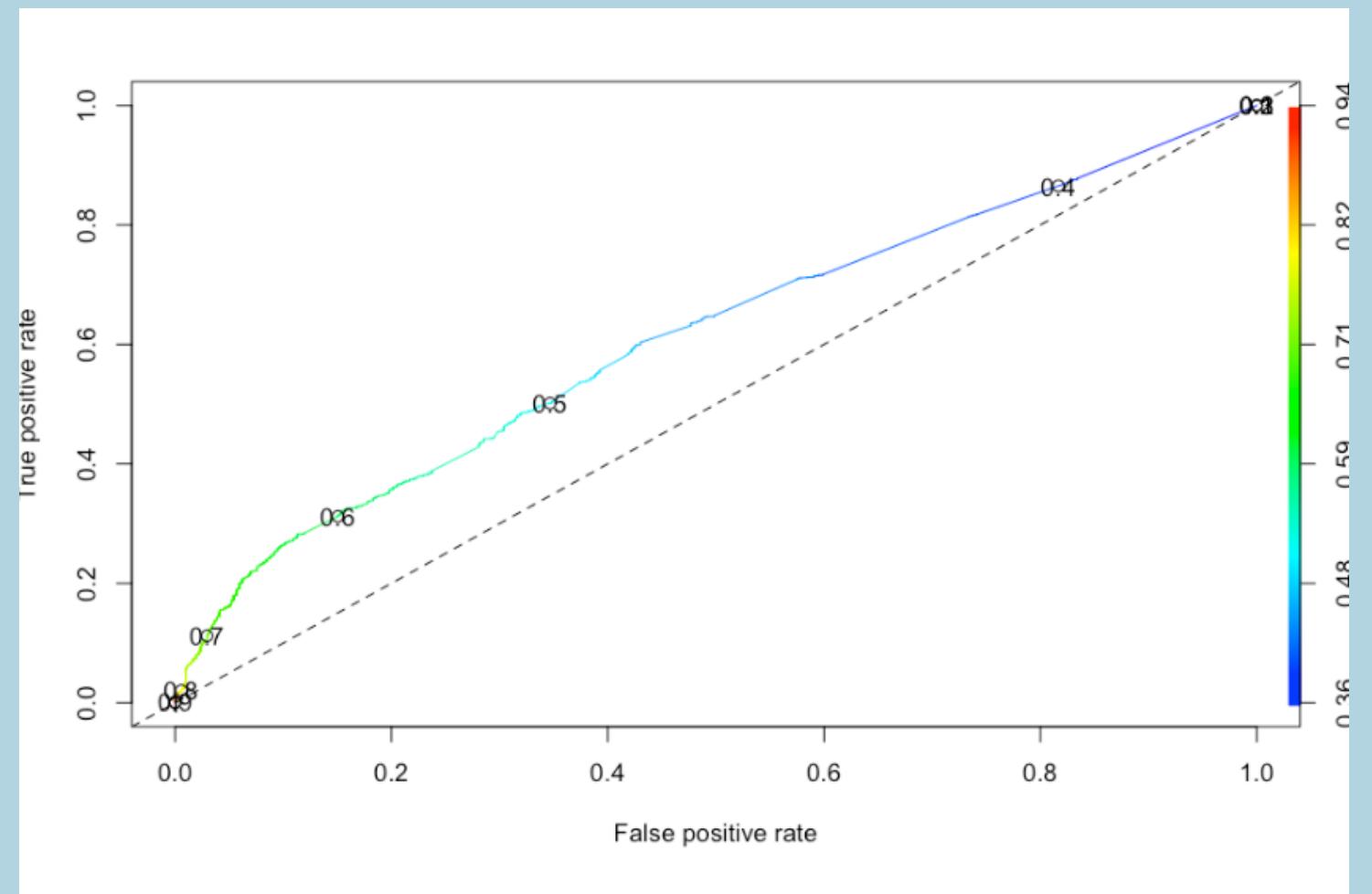
(Dispersion parameter for binomial family taken to be 1)

Null deviance: 23472 on 16931 degrees of freedom
Residual deviance: 22503 on 16926 degrees of freedom
AIC: 22515

Number of Fisher Scoring iterations: 4

*DETERMINED THROUGH COMPARASION OF AIC & VALIDATION ON TEST SET

Roc Chart



AUC: .6108

Confusion Matrix

Accuracy : 0.5788
95% CI : (0.5673, 0.5901)
No Information Rate : 0.5775
P-Value [Acc > NIR] : 0.4202

Kappa : 0.1554

McNemar's Test P-Value : <2e-16

Sensitivity : 0.5753
Specificity : 0.5835
Pos Pred Value : 0.6537
Neg Pred Value : 0.5013
Precision : 0.6537
Recall : 0.5753
F1 : 0.6120
Prevalence : 0.5775
Detection Rate : 0.3322
Detection Prevalence : 0.5082
Balanced Accuracy : 0.5794

'Positive' Class : closed

ANALYSIS

TREES

BAGGING

BOOSTING

RANDOM FORESTS

USING A SINGLE TREE

Predictors Used:

01
**TOTAL
INVESTMENT**

02
**SEED
CAPITAL**

03
**VENTURE
CAPITAL**

04
**FUNDING
YEARS**

05
**FUNDING
ROUNDS**

06
**MARKET
TYPE**

Number of terminal nodes : 72

Misclassification error rate: 0.3314

Accuracy : 66.93%

Confusion Matrix and Statistics

		Reference	
		Prediction	
		0	1
0	2879	1610	
1	790	1978	

Accuracy : 0.6693
95% CI : (0.6583, 0.6801)

No Information Rate : 0.5056
P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.3368

McNemar's Test P-value : < 2.2e-16

Sensitivity : 0.7847
Specificity : 0.5513
Pos Pred Value : 0.6413
Neg Pred Value : 0.7146
Precision : 0.6413
Recall : 0.7847
F1 : 0.7058
Prevalence : 0.5056
Detection Rate : 0.3967
Detection Prevalence : 0.6186
Balanced Accuracy : 0.6680

'Positive' class : 0

Confusion Matrix and statistics

		Reference	
		0	1
Prediction	0	3577	900
	1	92	2688

Accuracy : 0.8633

95% CI : (0.8552, 0.8711)

No Information Rate : 0.5056

P-value [Acc > NIR] : < 2.2e-16

Kappa : 0.7259

McNemar's Test P-value : < 2.2e-16

Sensitivity : 0.9749

Specificity : 0.7492

Pos Pred Value : 0.7990

Neg Pred Value : 0.9669

Precision : 0.7990

Recall : 0.9749

F1 : 0.8782

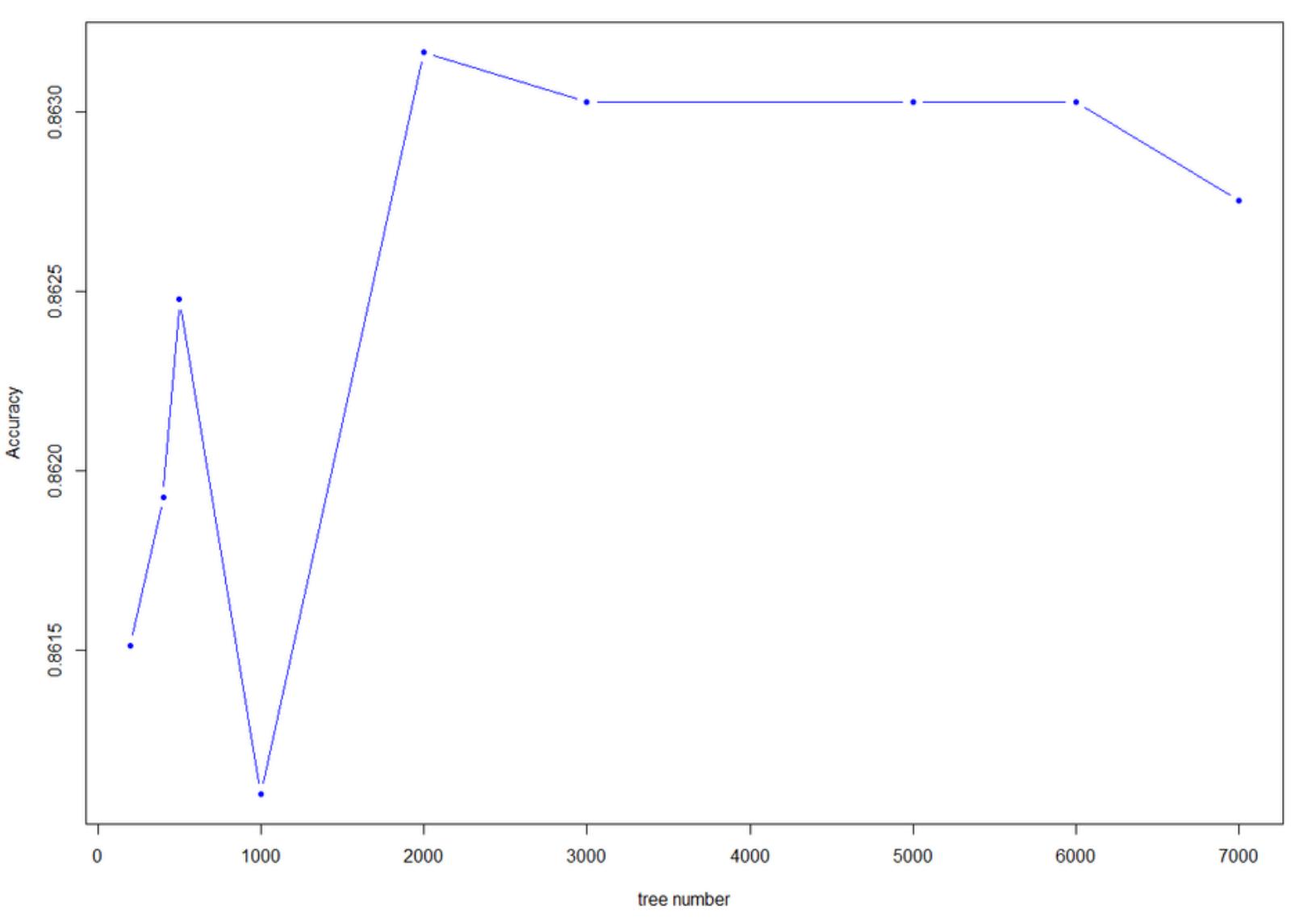
Prevalence : 0.5056

Detection Rate : 0.4929

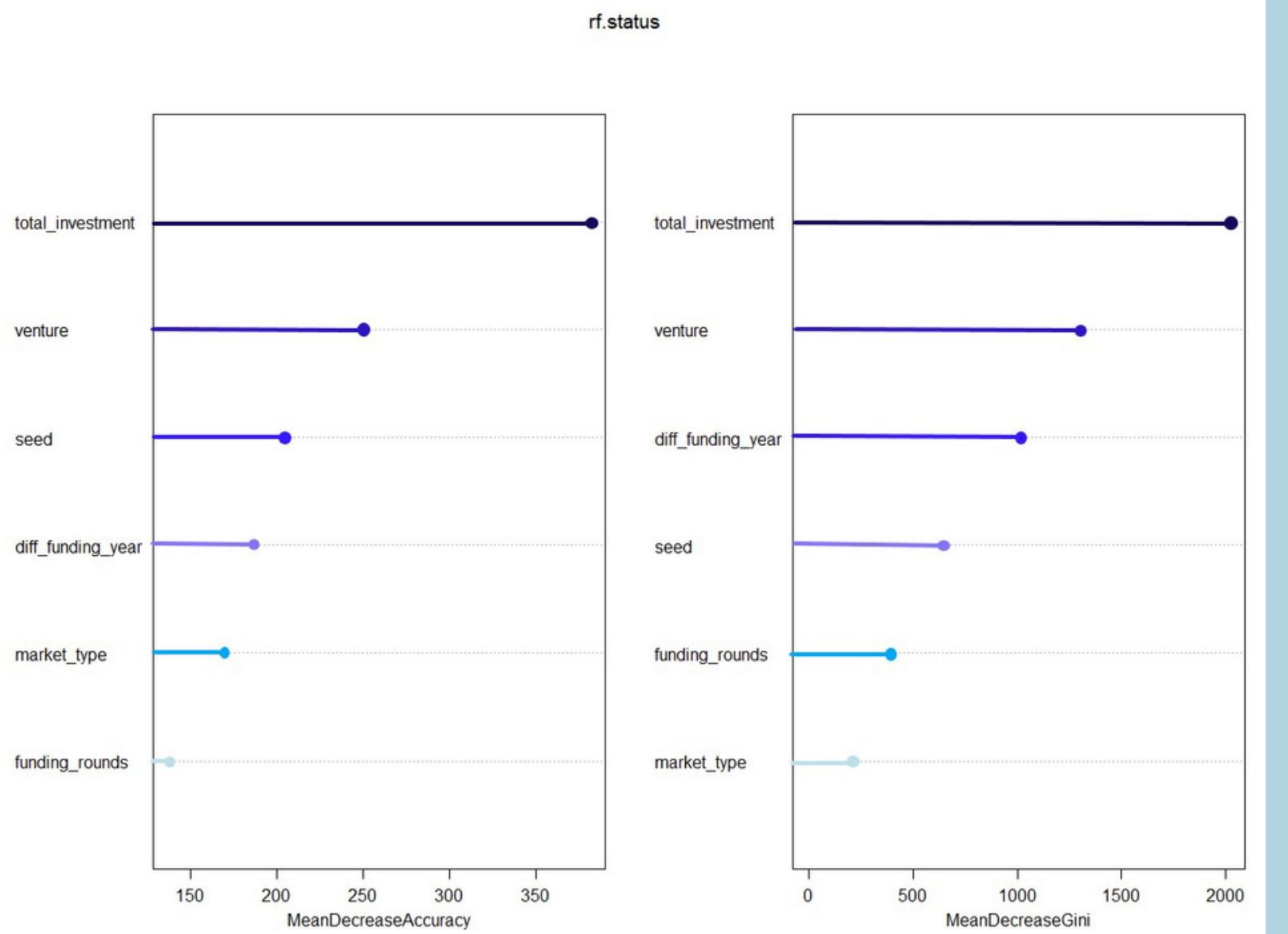
Detection Prevalence : 0.6169

Balanced Accuracy : 0.8620

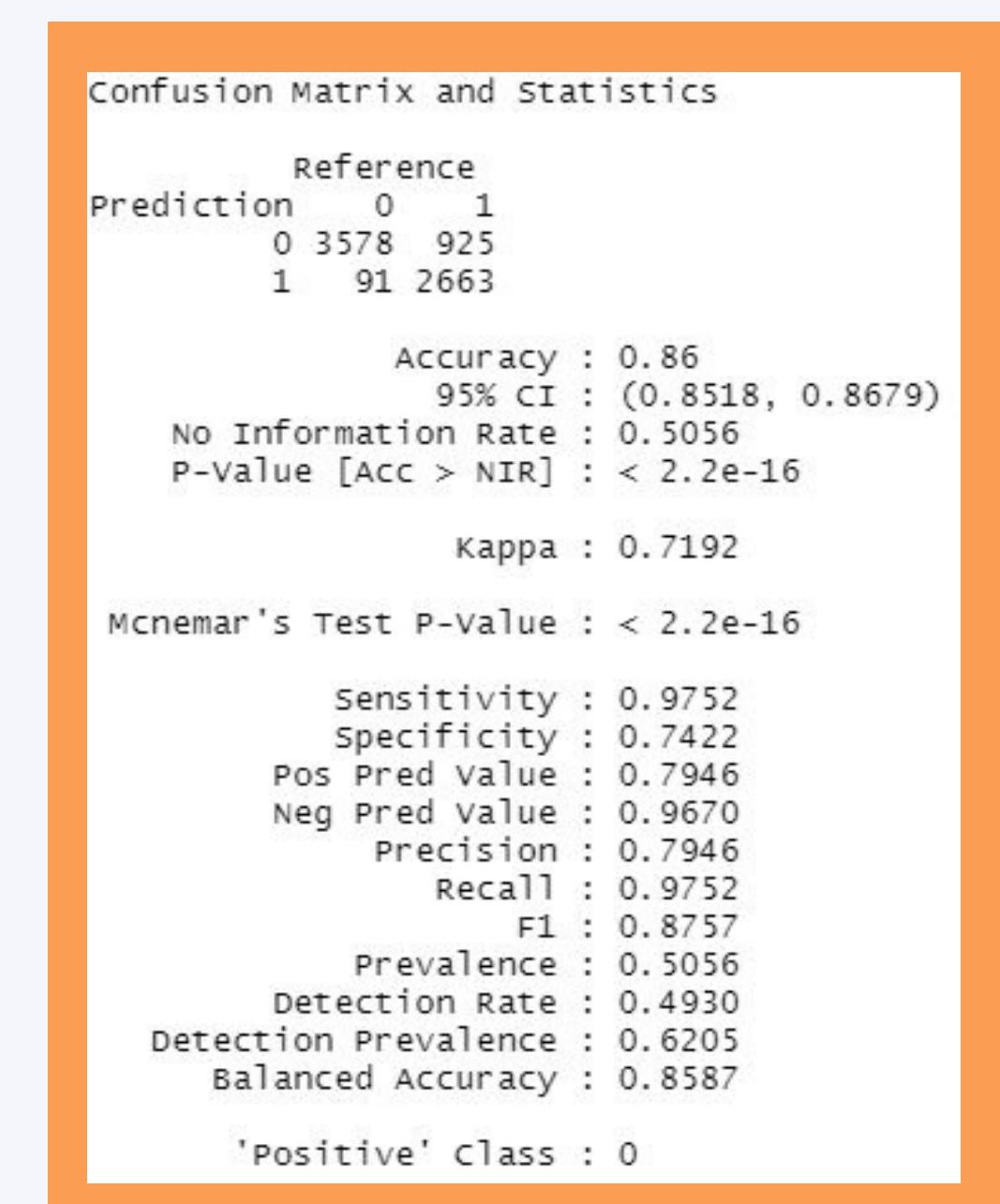
'Positive' class : 0



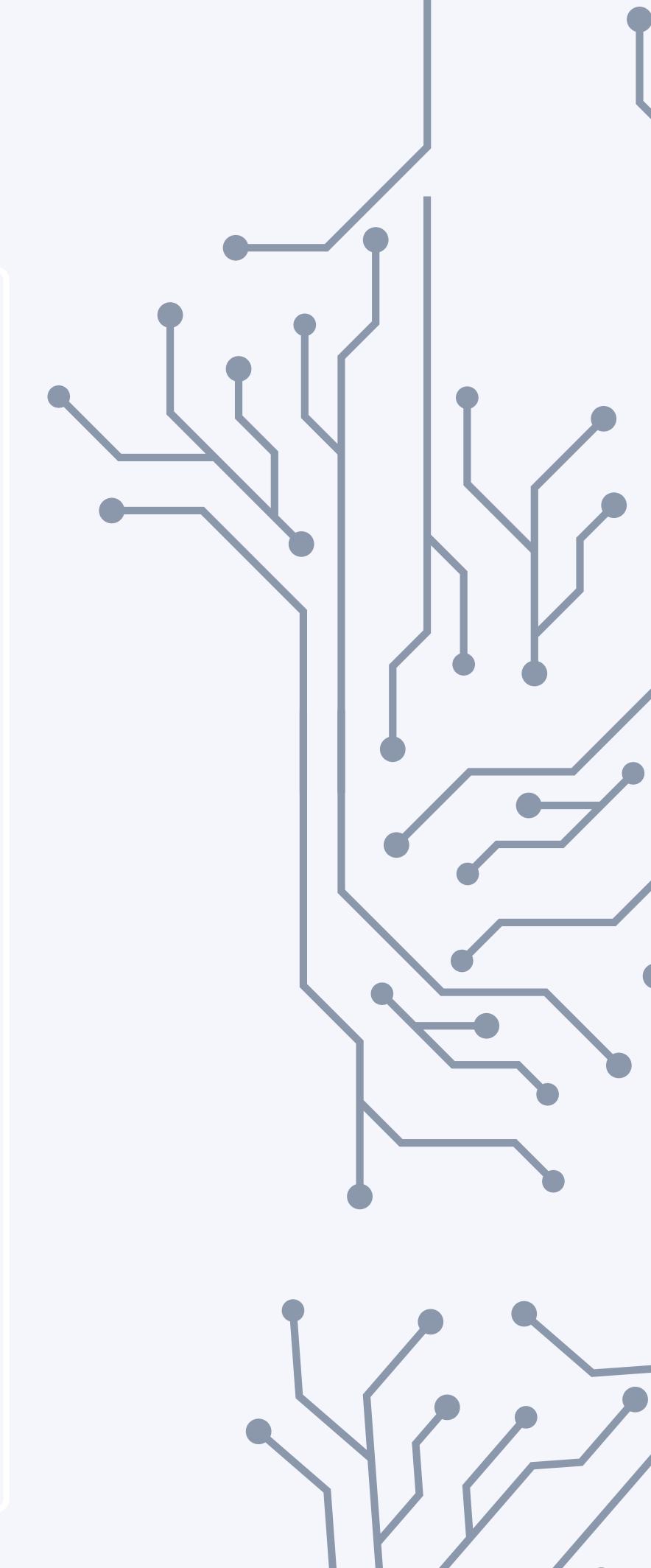
RANDOM FOREST



01 *Subset of predictors*
Resulted in the reveal of 2 most important variables



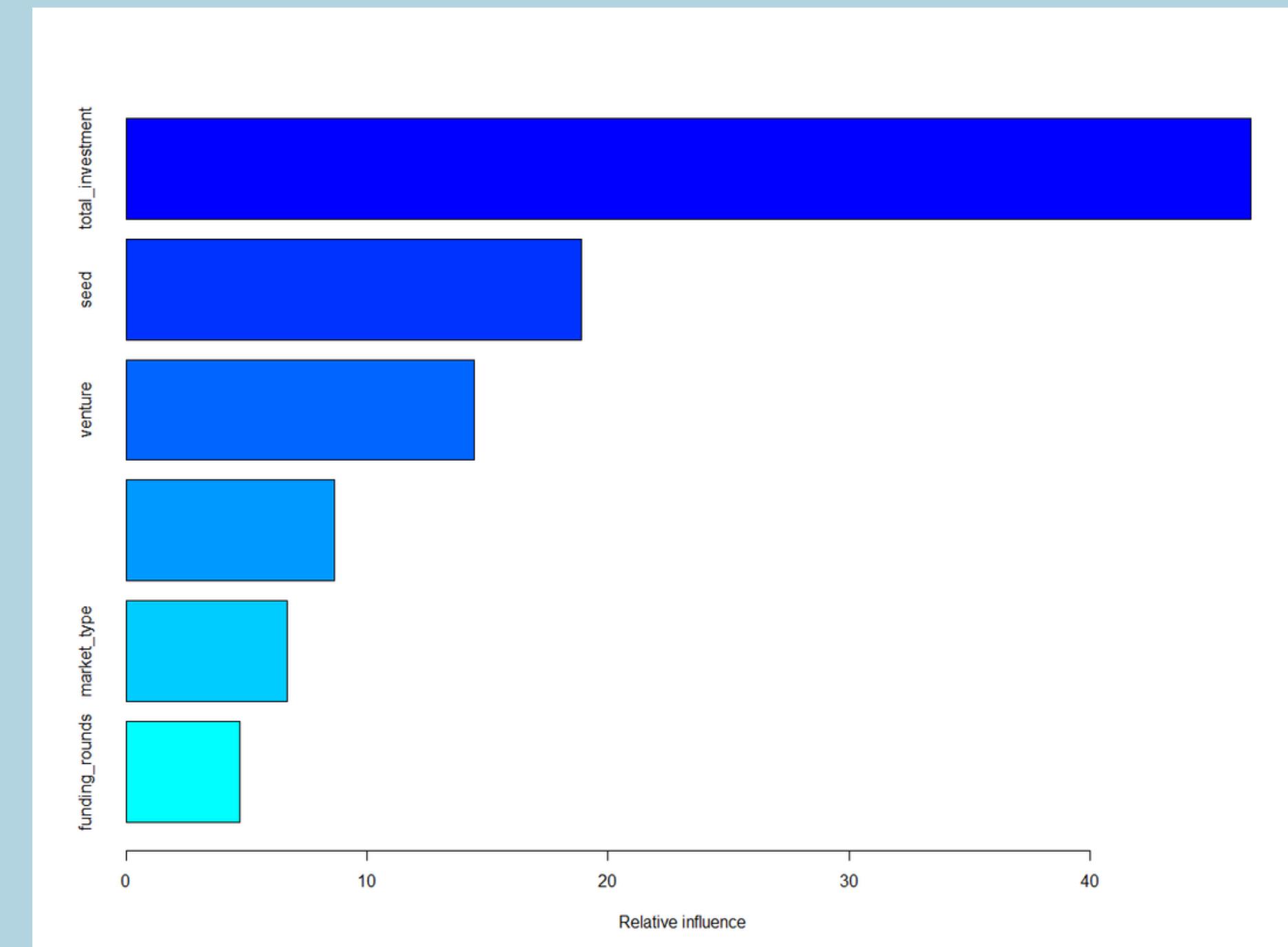
02 *Confusion Matrix*
This confirms our results from previous models



Different combinations of B, d, λ

Tdepth (d)	Ntree (B)	Lambda	Accuracy
3	100	0.2	68.97
3	1000	0.2	81.4
3	3000	0.2	84.26
3	4000	0.2	84.5
3	5000	0.2	83.04
4	1000	0.2	82.39
4	1000	0.01	67.23
4	4000	0.01	75.53
10	1000	0.01	75.03
10	3000	0.01	71.62

BOOSTING



RESULTS



**TOTAL
INVESTMENT**



**VENTURE
CAPITAL**



**FUNDING
YEARS**

THANK YOU!

Any questions?