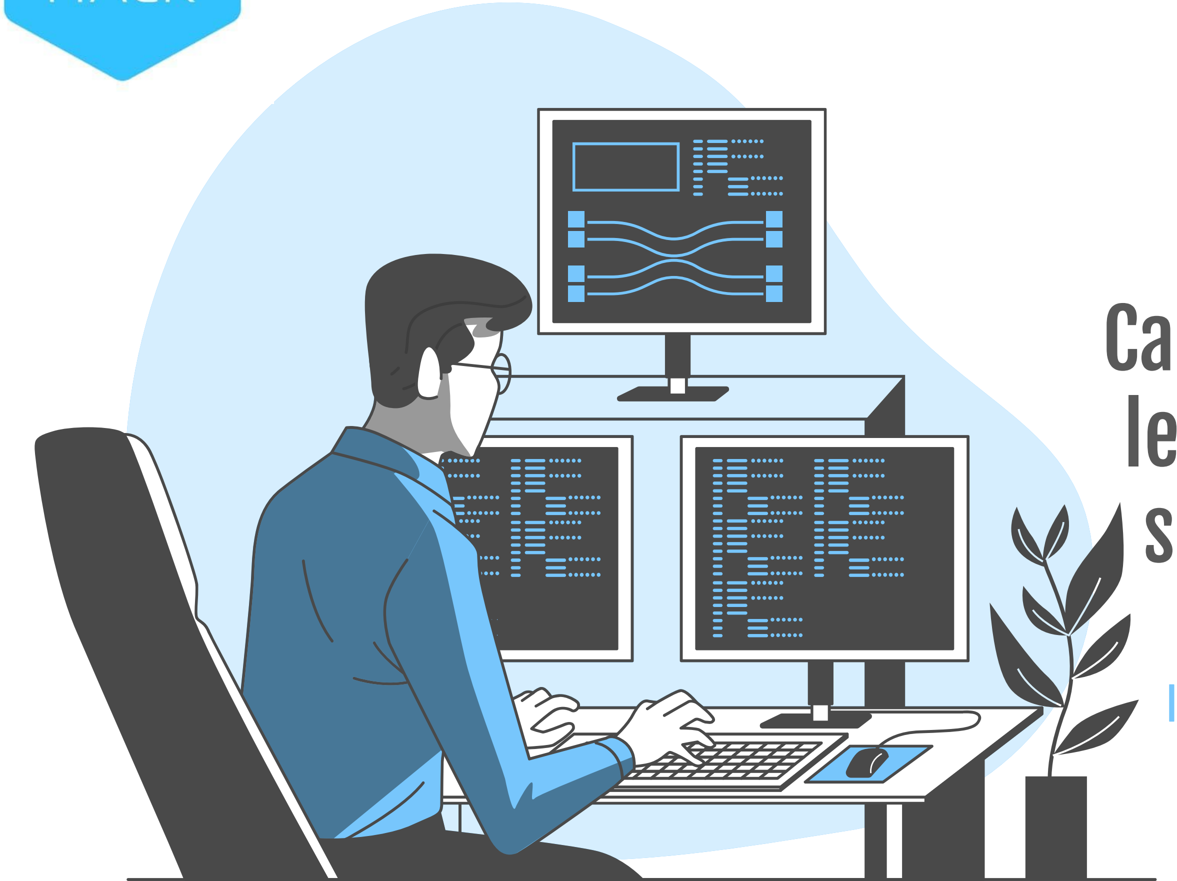


Can we use machine learning to predict startup success ?

Kyrian Bourgi

Ironhack Final Presentation

Data Analytics

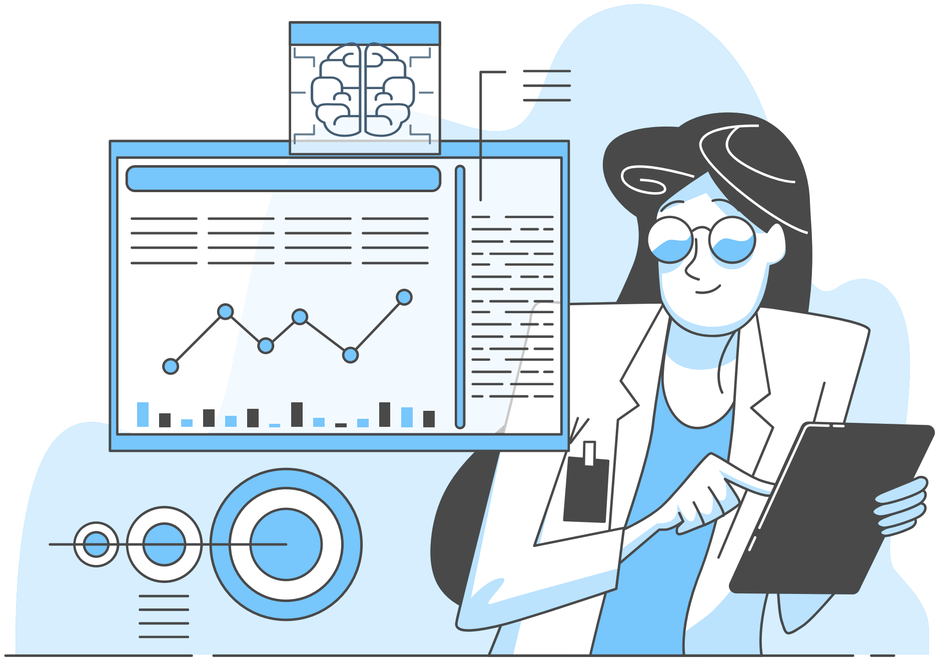
Table of Contents

01 Problem Definition

02 Data Overview

03 Model Results

04 Conclusion



Problem Definition

1. How do we measure **Success?**

2. Which numerical feature can we base our predictions on ?

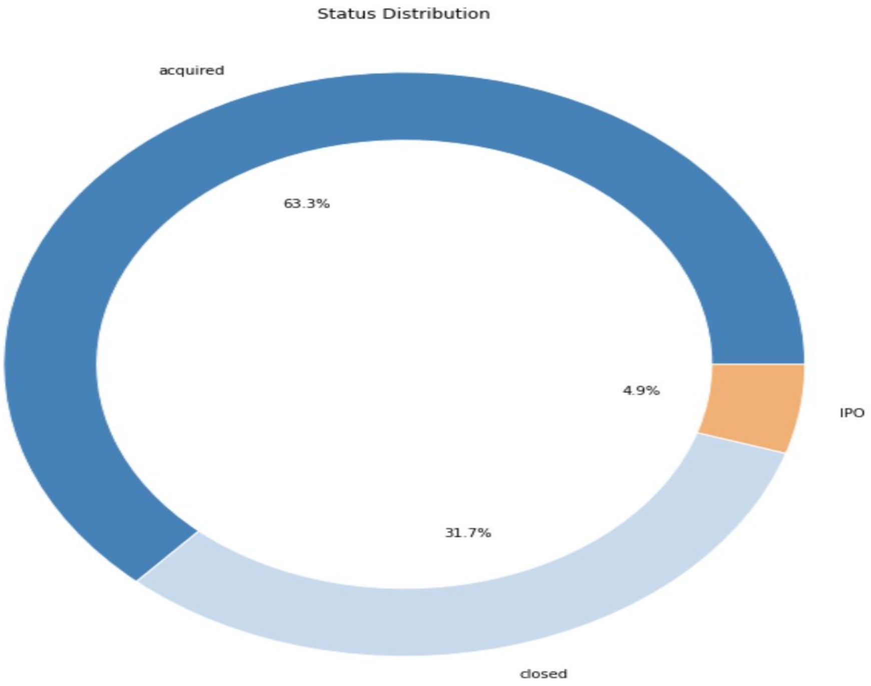
3. Compare model accuracies (Logistic Regression & Random Forest)

4. What financing option are more successful in predicting the success of a company ?

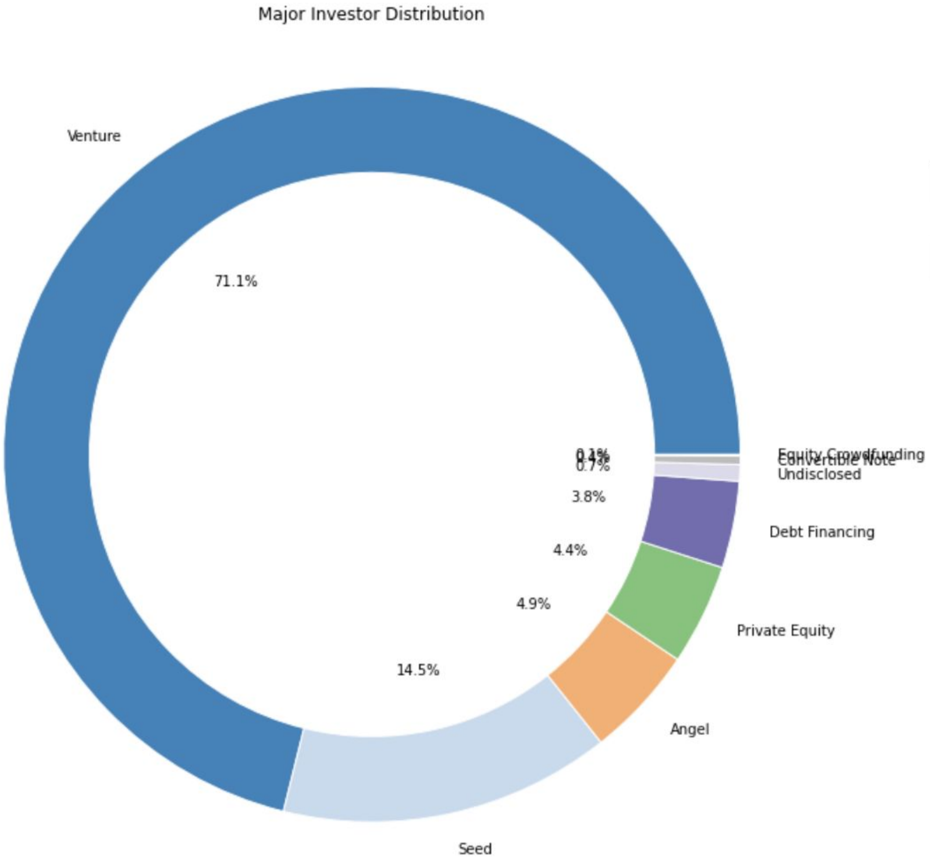
Analysis

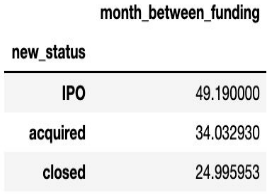
Outreach

Status Distribution



- As we can see we have 4.9% of our companies that are our "Successful" / IPO
- 63% that we can classify as "Successful Acquisition" / Acquired
- 31.7% that we can classify as "Failed" / Closed

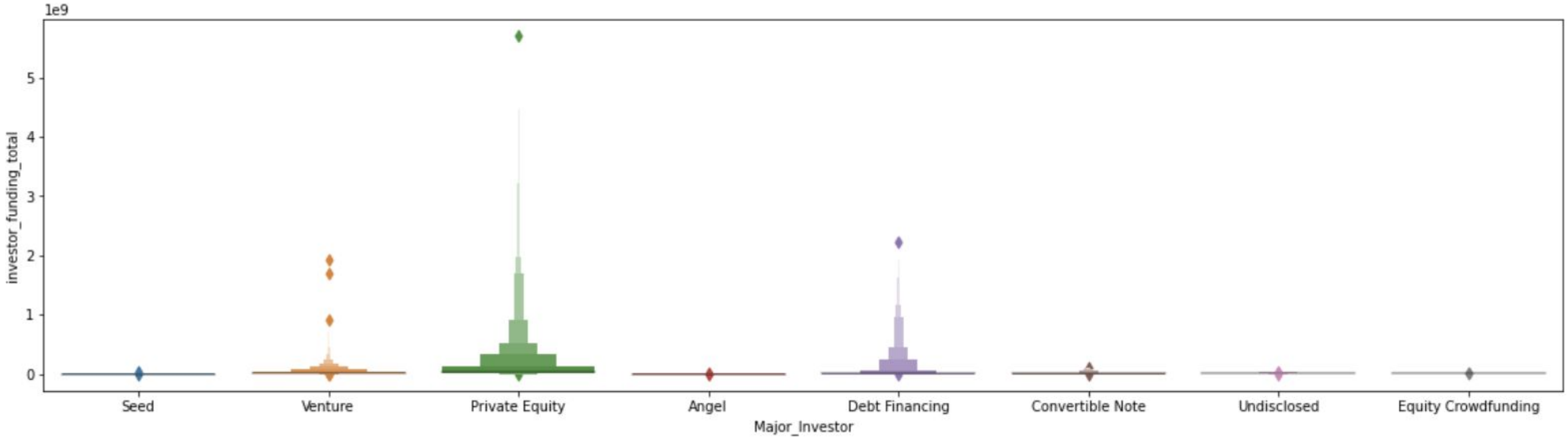




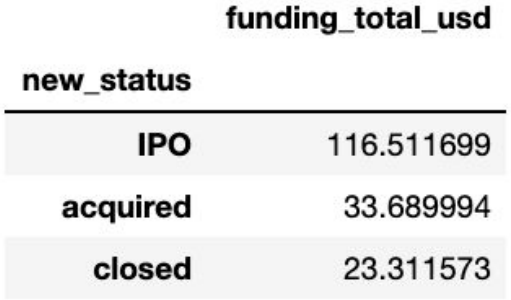
	month_between_funding
new_status	
IPO	49.190000
acquired	34.032930
closed	24.995953

- We have on average 49 months between the first and the last round of funding for IPO
- 34 months between between the first and the last round of funding for acquired
- 25 months between the first and the last round of funding for closed

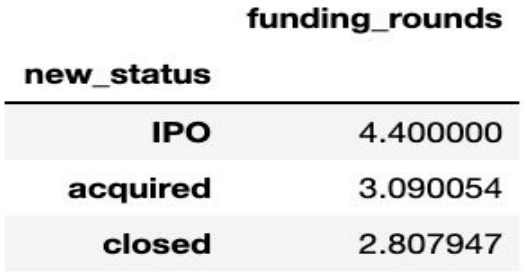
Distribution of the Major Financing Options



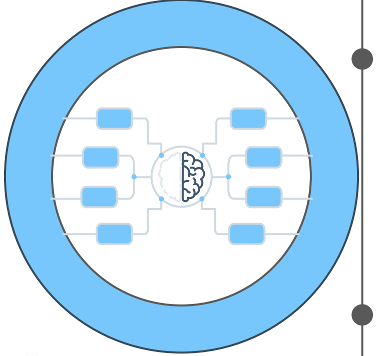
We can observe that most of the total funding comes from Private Equity, then Venture Capital and thirdly Debt Financing

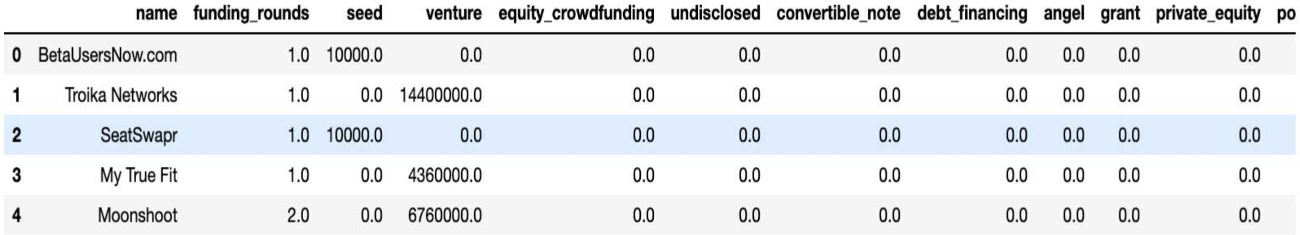


	funding_total_usd
new_status	
IPO	116.511699
acquired	33.689994
closed	23.311573



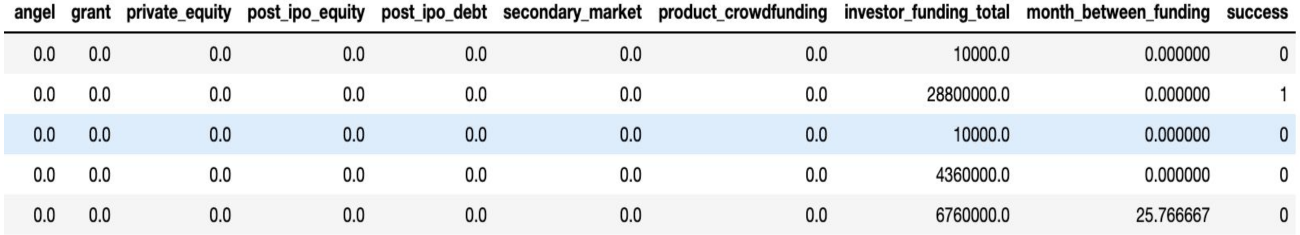
	funding_rounds
new_status	
IPO	4.400000
acquired	3.090054
closed	2.807947





	name	funding_rounds	seed	venture	equity_crowdfunding	undisclosed	convertible_note	debt_financing	angel	grant	private_equity	po
0	BetaUsersNow.com	1.0	10000.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
1	Troika Networks	1.0	0.0	14400000.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
2	SeatSwapr	1.0	10000.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
3	My True Fit	1.0	0.0	4360000.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
4	Moonshoot	2.0	0.0	6760000.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	

1968 rows × 18 columns

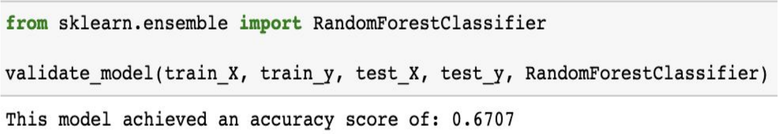


angel	grant	private_equity	post_ipo_equity	post_ipo_debt	secondary_market	product_crowdfunding	investor_funding_total	month_between_funding	success
0.0	0.0	0.0	0.0	0.0	0.0	0.0	10000.0	0.000000	0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	28800000.0	0.000000	1
0.0	0.0	0.0	0.0	0.0	0.0	0.0	10000.0	0.000000	0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	4360000.0	0.000000	0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	6760000.0	25.766667	0

How would we predict success ?

Machine Learning

Random Forest

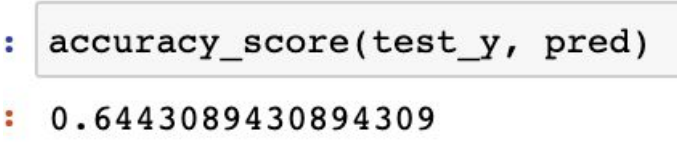


```
from sklearn.ensemble import RandomForestClassifier

validate_model(train_X, train_y, test_X, test_y, RandomForestClassifier)
```

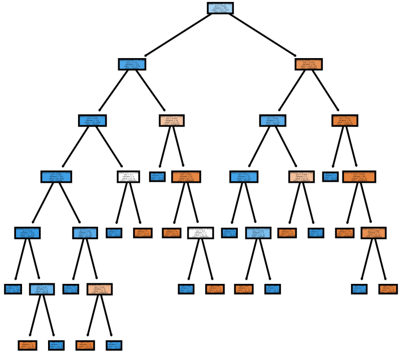
```
This model achieved an accuracy score of: 0.6707
```

Logistic Regression

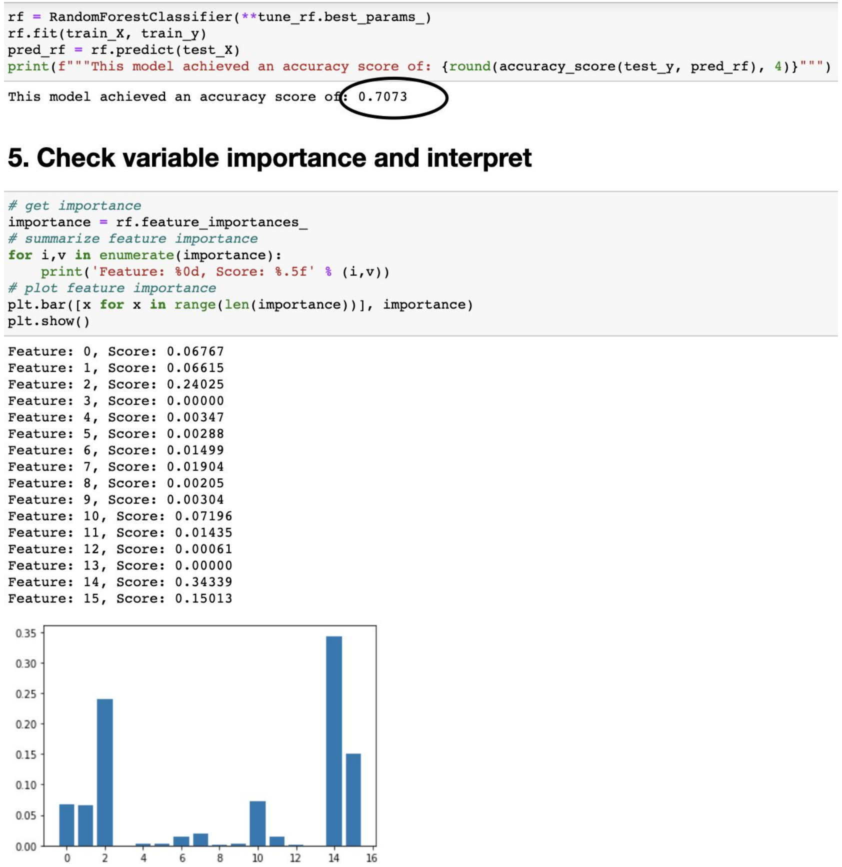


```
: accuracy_score(test_y, pred)

: 0.6443089430894309
```



Results



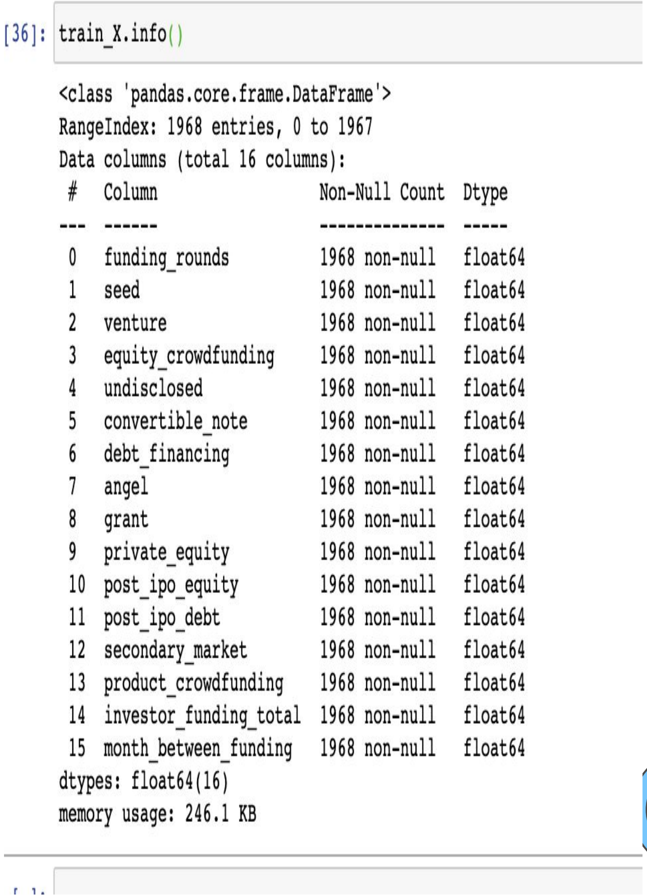
```
rf = RandomForestClassifier(**tune_rf.best_params_)
rf.fit(train_X, train_y)
pred_rf = rf.predict(test_X)
print(f"""This model achieved an accuracy score of: {round(accuracy_score(test_y, pred_rf), 4)}""")
```

```
This model achieved an accuracy score of: 0.7073
```

5. Check variable importance and interpret

```
# get importance
importance = rf.feature_importances_
# summarize feature importance
for i,v in enumerate(importance):
    print('Feature: %0d, Score: %.5f' % (i,v))
# plot feature importance
plt.bar([x for x in range(len(importance))], importance)
plt.show()
```

```
Feature: 0, Score: 0.06767
Feature: 1, Score: 0.06615
Feature: 2, Score: 0.24025
Feature: 3, Score: 0.00000
Feature: 4, Score: 0.00347
Feature: 5, Score: 0.00288
Feature: 6, Score: 0.01499
Feature: 7, Score: 0.01904
Feature: 8, Score: 0.00205
Feature: 9, Score: 0.00304
Feature: 10, Score: 0.07196
Feature: 11, Score: 0.01435
Feature: 12, Score: 0.00061
Feature: 13, Score: 0.00000
Feature: 14, Score: 0.34339
Feature: 15, Score: 0.15013
```

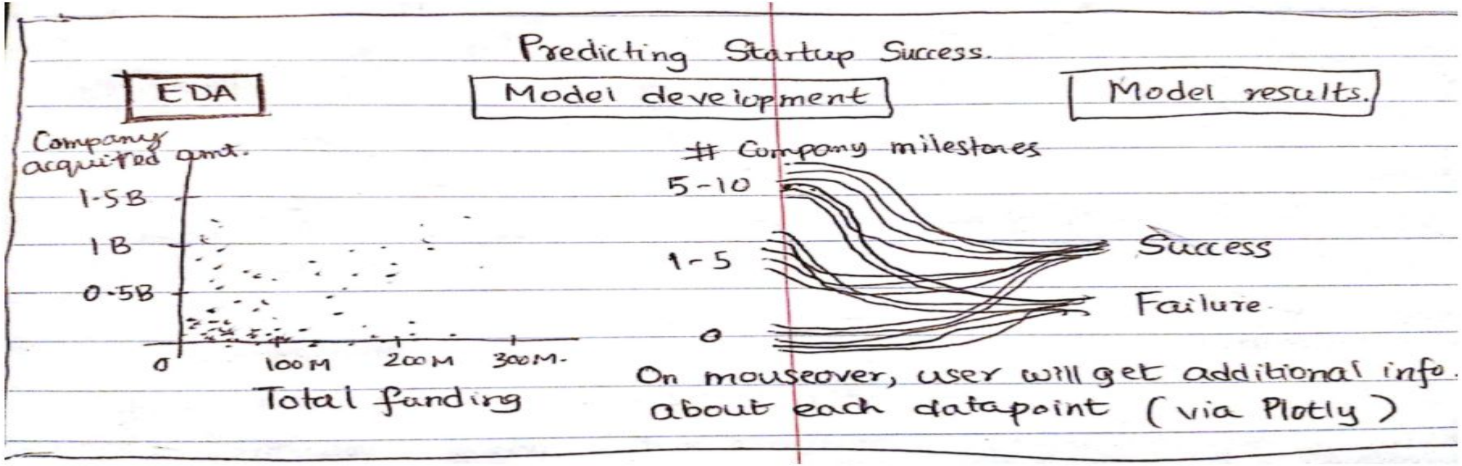


```
[36]: train_X.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1968 entries, 0 to 1967
Data columns (total 16 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   funding_rounds        1968 non-null   float64
 1   seed                  1968 non-null   float64
 2   venture               1968 non-null   float64
 3   equity_crowdfunding   1968 non-null   float64
 4   undisclosed           1968 non-null   float64
 5   convertible_note      1968 non-null   float64
 6   debt_financing        1968 non-null   float64
 7   angel                 1968 non-null   float64
 8   grant                 1968 non-null   float64
 9   private_equity        1968 non-null   float64
 10  post_ipo_equity       1968 non-null   float64
 11  post_ipo_debt         1968 non-null   float64
 12  secondary_market      1968 non-null   float64
 13  product_crowdfunding  1968 non-null   float64
 14  investor_funding_total 1968 non-null  float64
 15  month_between_funding 1968 non-null   float64
dtypes: float64(16)
memory usage: 246.1 KB
```

```
[ ]:
```

Model criticism & Open discussion



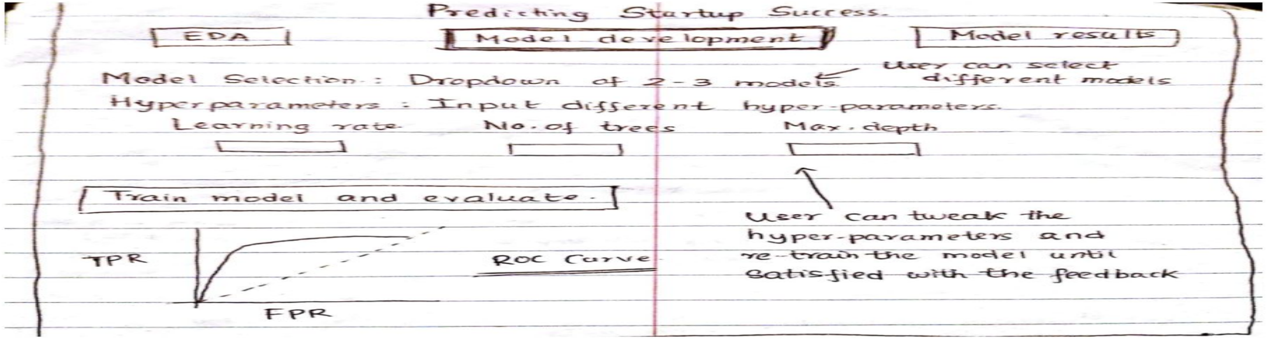
Predicting Startup Success.

EDA

Model development

Model results.

On mouseover, user will get additional info. about each datapoint (via Plotly)



Predicting Startup Success.

EDA

Model development

Model results

Model Selection: Dropdown of 2-3 models

User can select different models

Hyperparameters: Input different hyper-parameters.

Learning rate □

No. of trees □

Max. depth □

User can tweak the hyper-parameters and re-train the model until satisfied with the feedback

Train model and evaluate.