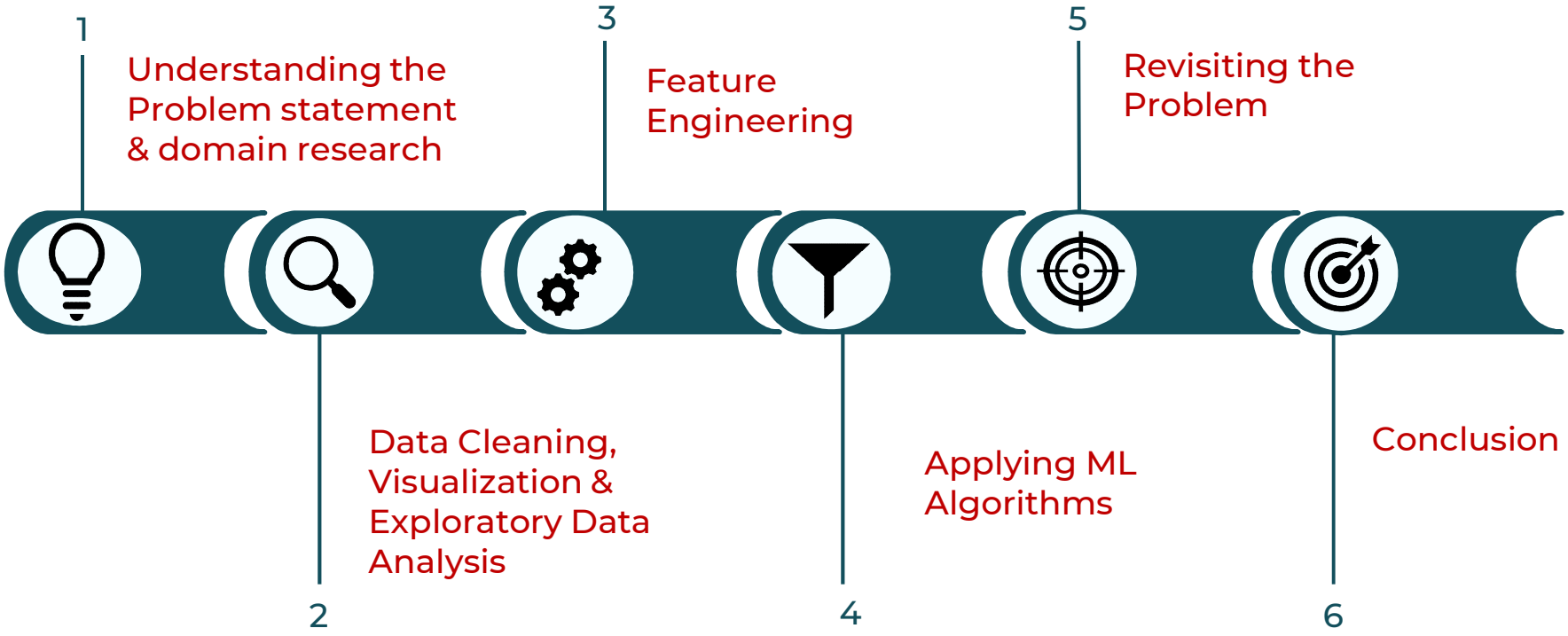
Capstone Project – 4

Start-up Funding Prediction

Is this startup ready for funding?Let’s find out!



Apprehending the goal!

There has been a staggering growth in investments in young age startups in the last 5 years. A lot of big VC firms are increasingly getting interested in the startup funding space. We are given a task to predict whether a startup will get a funding in the next three months using app traction data and startup details.

Data Pallete

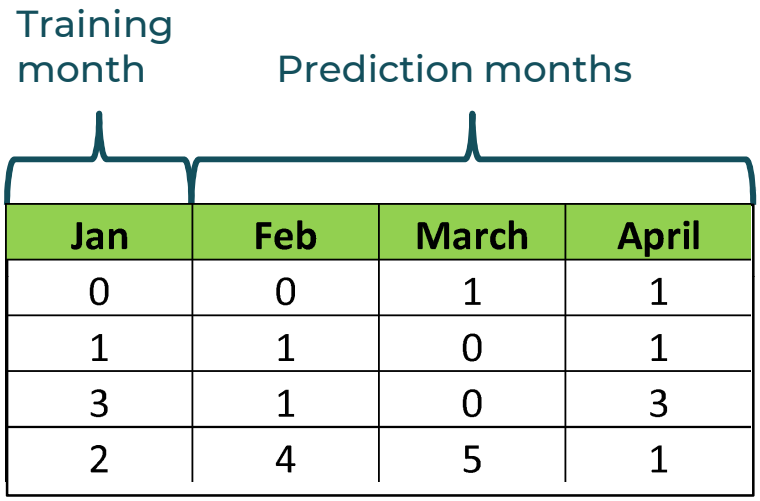
General Features

- UUID- (Unique Identifier)
- Month
- Application Category
- Avg Session Time
- Total Session Time
- Open Rate
- Reach
- Funding Ind

Features for funded Start-ups

- Business models
- City
- Company Stage
- Founded year
- Latest funded date
- Monthly Active Users
- Total Funding
- Uninstall Rate

Approaching the Problem



Jan	Feb	March	April
0	0	1	1
1	1	0	1
3	1	0	3
2	4	5	1

- Dropped the features which were only provided for funded startups.
- Taking entire dataset.
- As, we want to predict whether start-up will be funded in next 3 months, we took the last 3 months as prediction months and made the month before 3 months as th training month.

Data Cleaning

Initial Data size

(1502175, 20)

Removing Duplicates

No. of Duplicates: 89412

Removing features that were only present for Funded startups & Missing Values Imputation

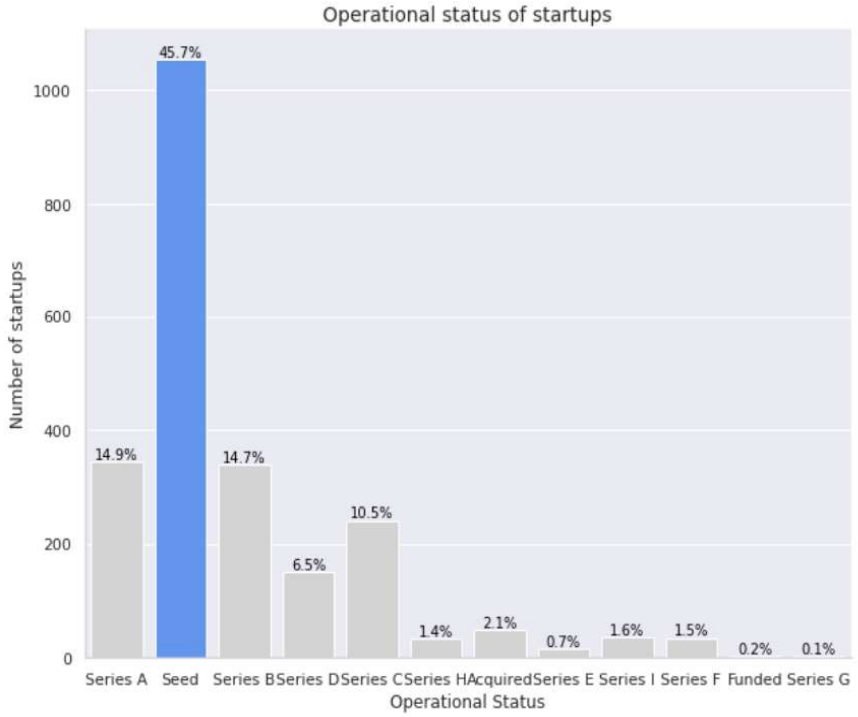
Imputing missing values by creating a class that would go for each UUID then by Category and apply back fill(limit-1) & forward fill(limit-1) and remaining with mean of that UUID and category.

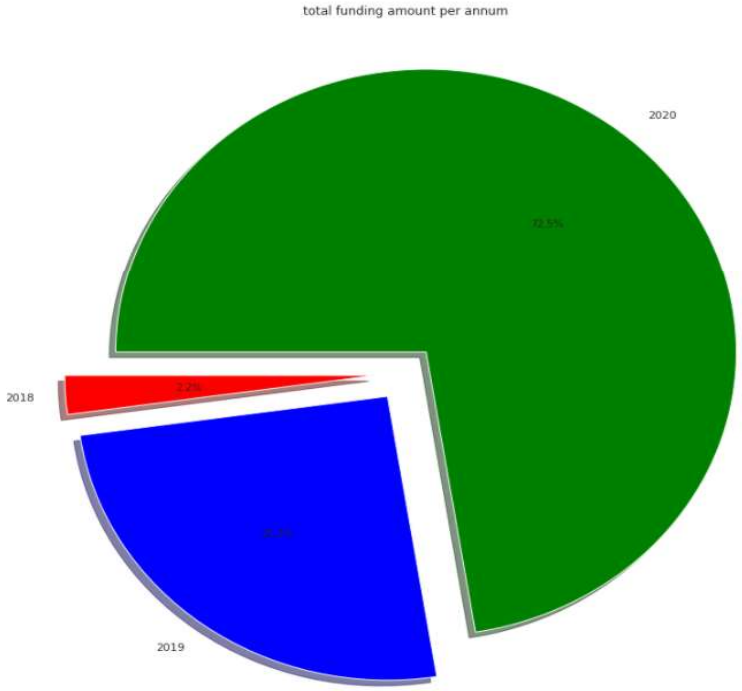
UUID with count less than 4

According to the approach discussed above, UUID with less that 4 months data is removed

UUID	Application Category	
x	Business	1
x	Business	1
x	Business	1
y	Entertainment	2
y	Entertainment	2
y	Entertainment	2

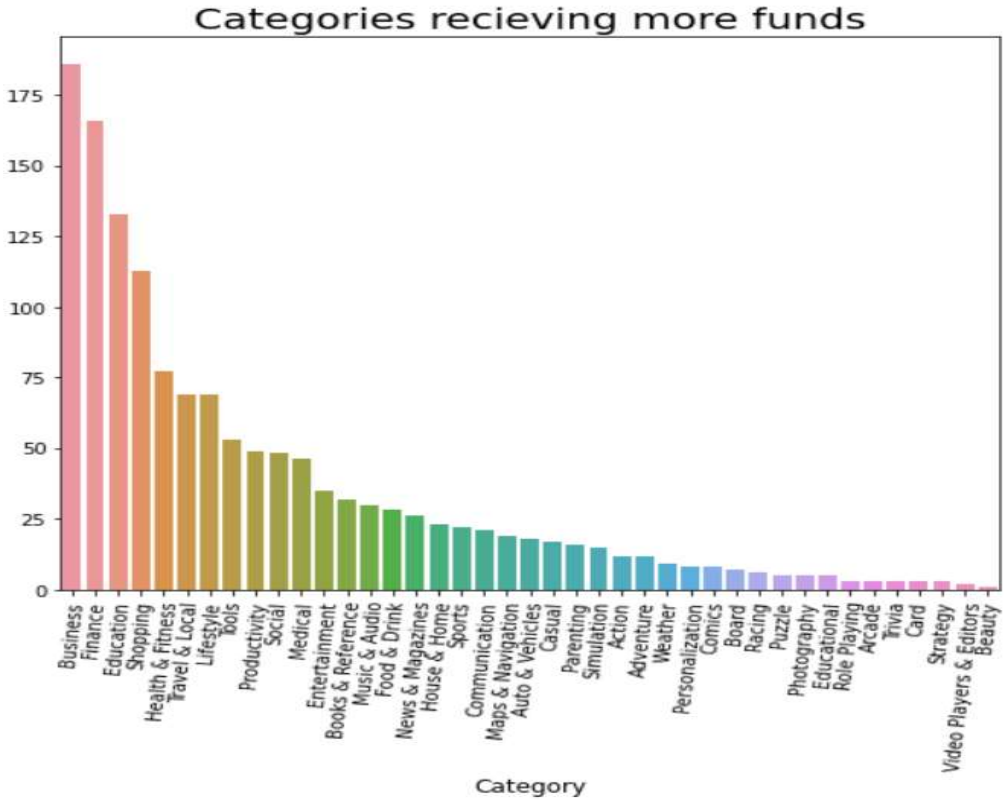
Exploratory Data Analysis

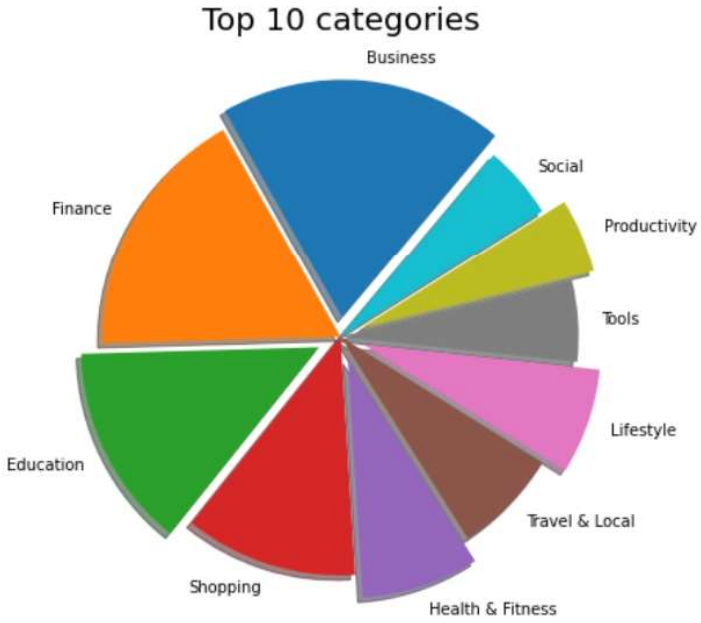




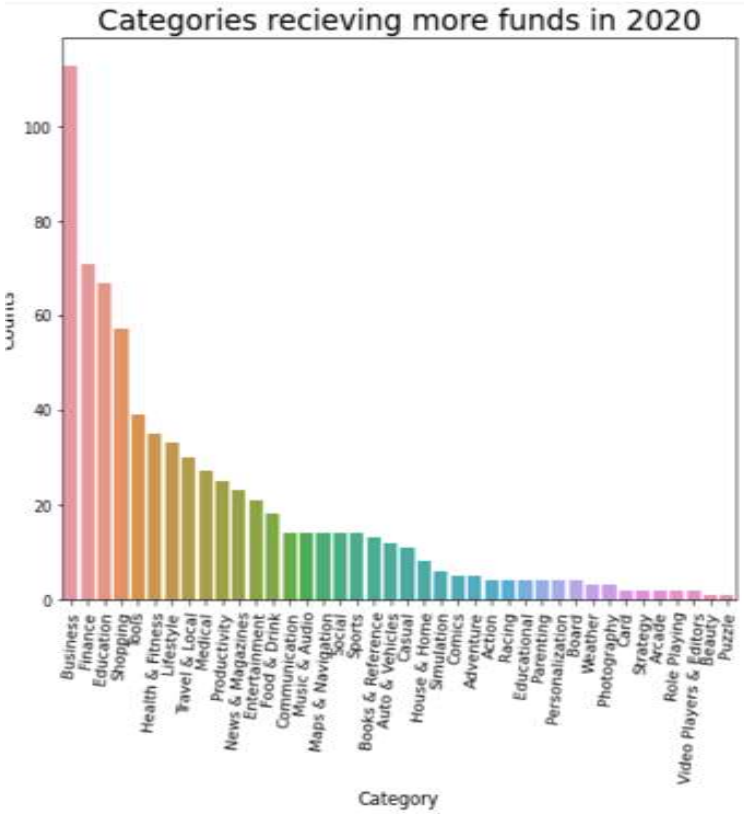
Exploratory Data Analysis

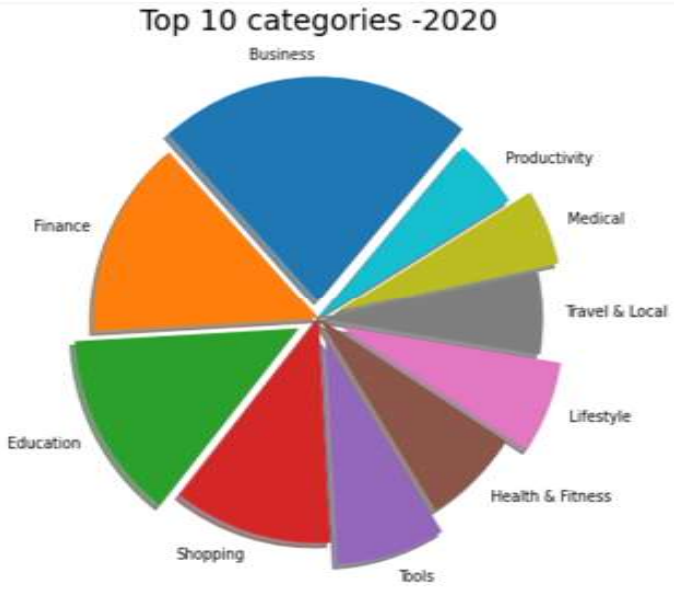
Exploratory Data Analysis



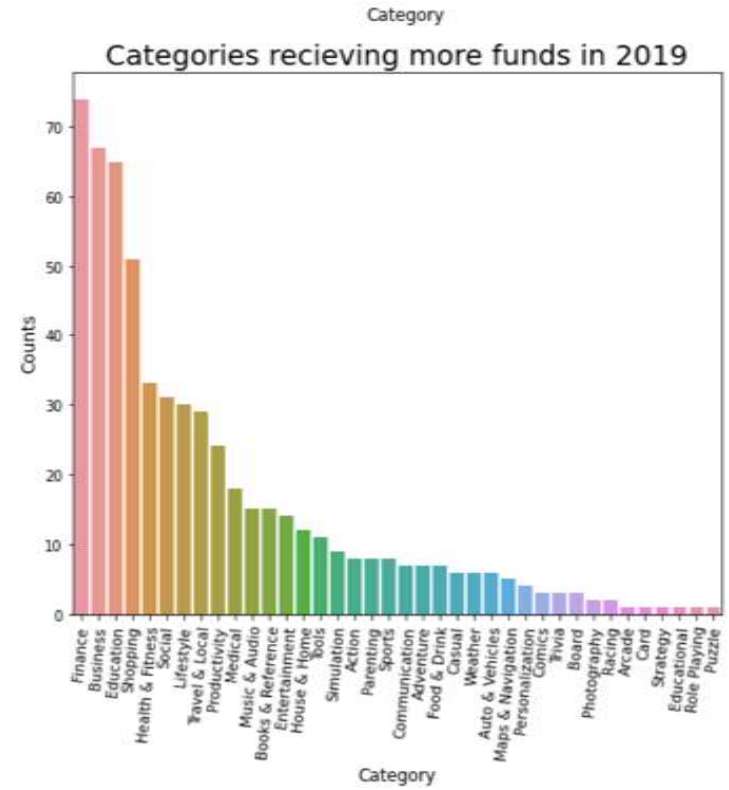


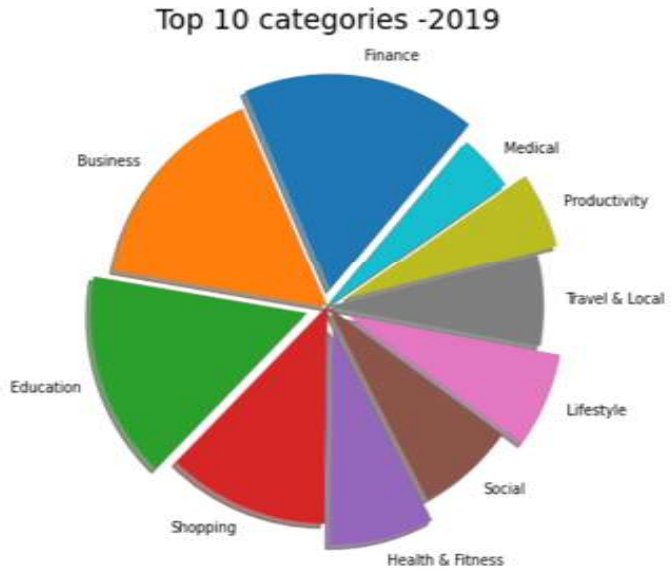
Exploratory Data Analysis



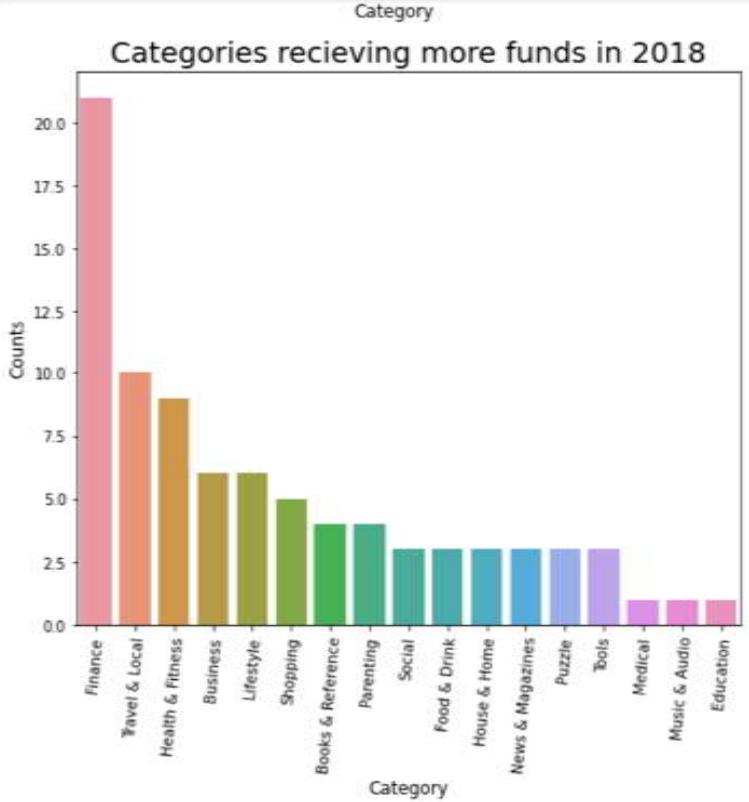


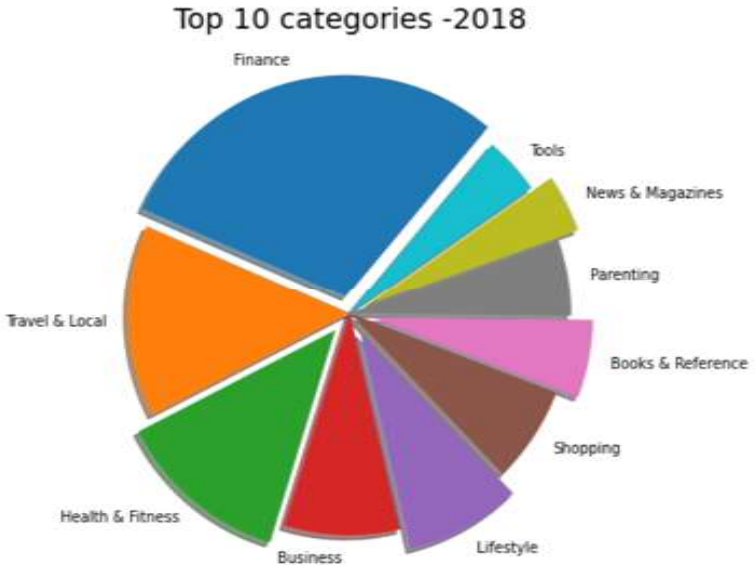
Exploratory Data Analysis



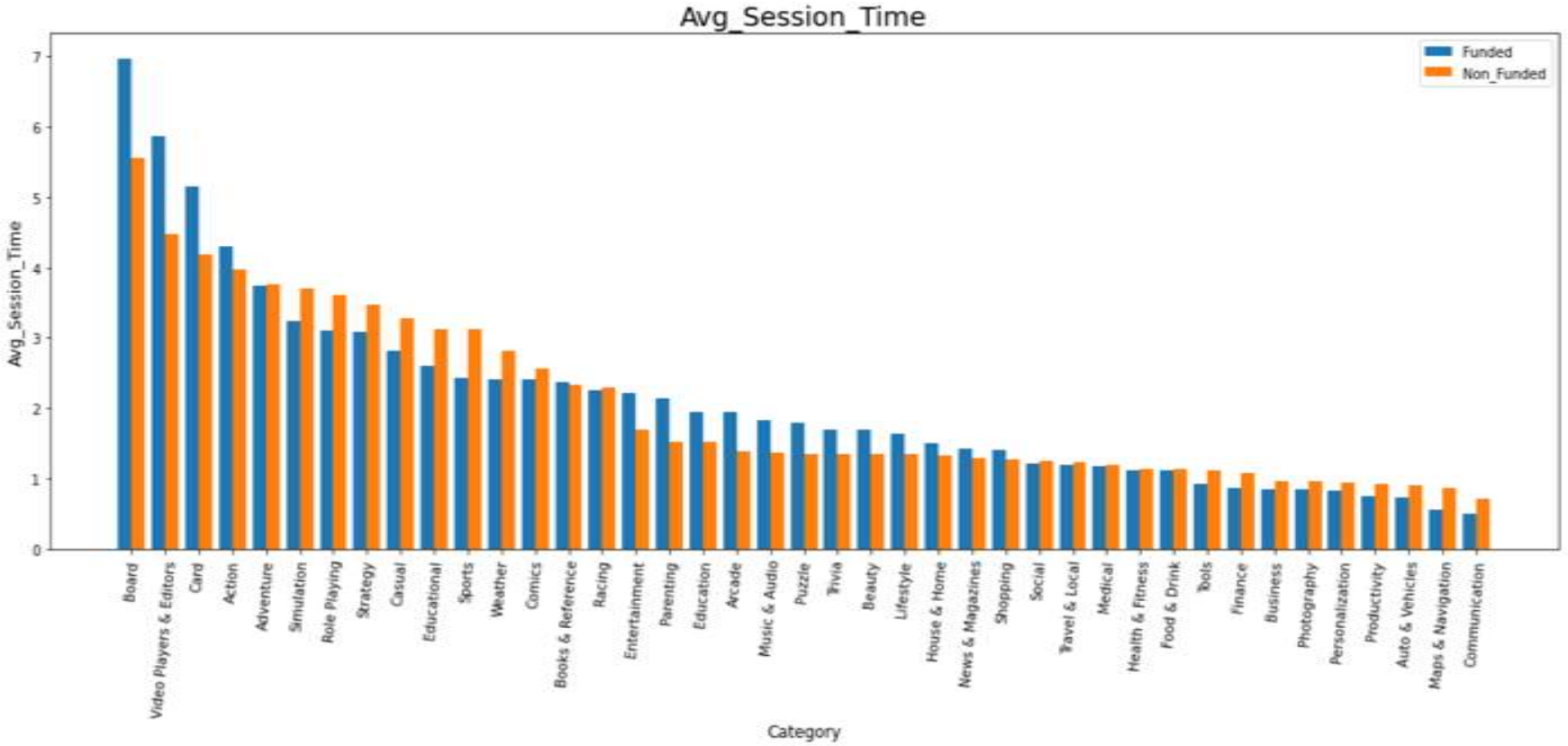


Exploratory Data Analysis

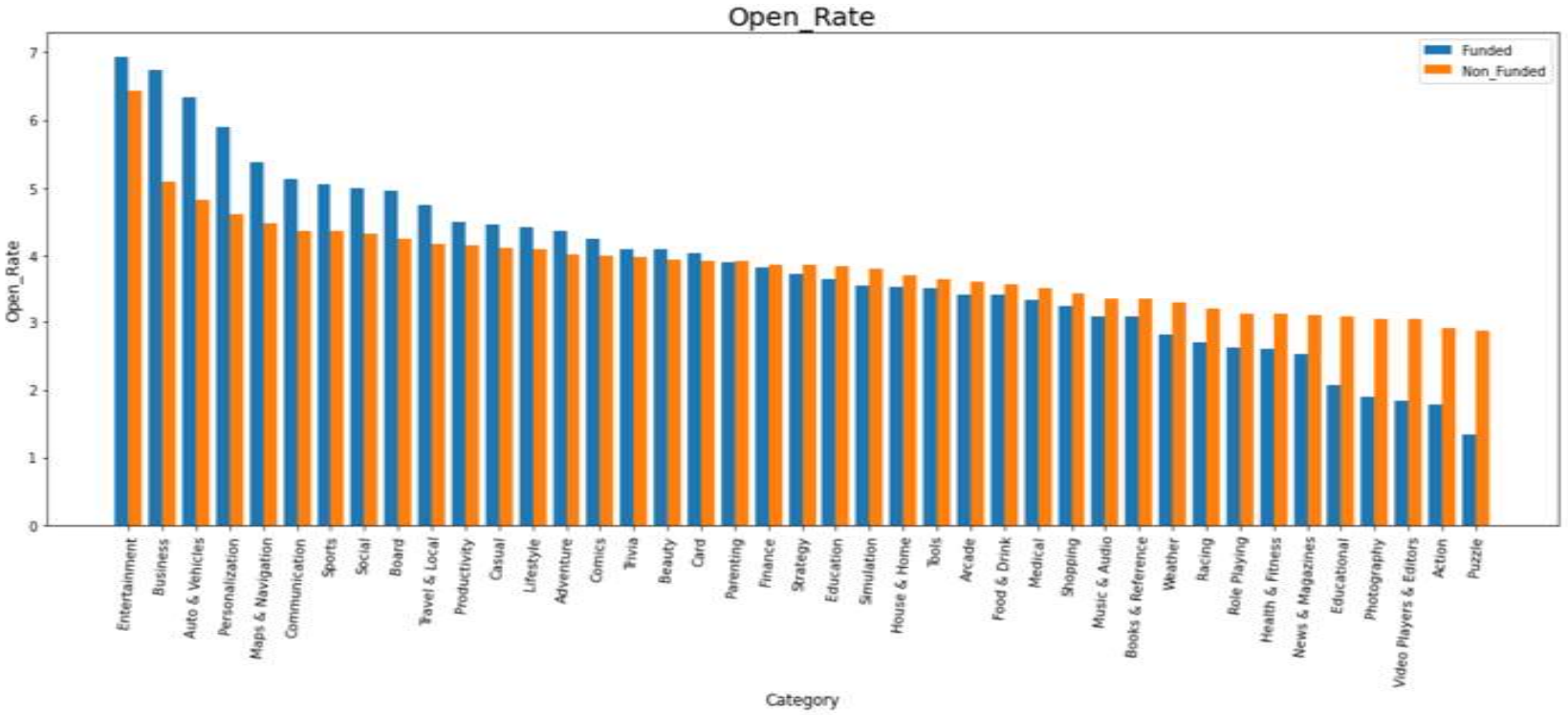




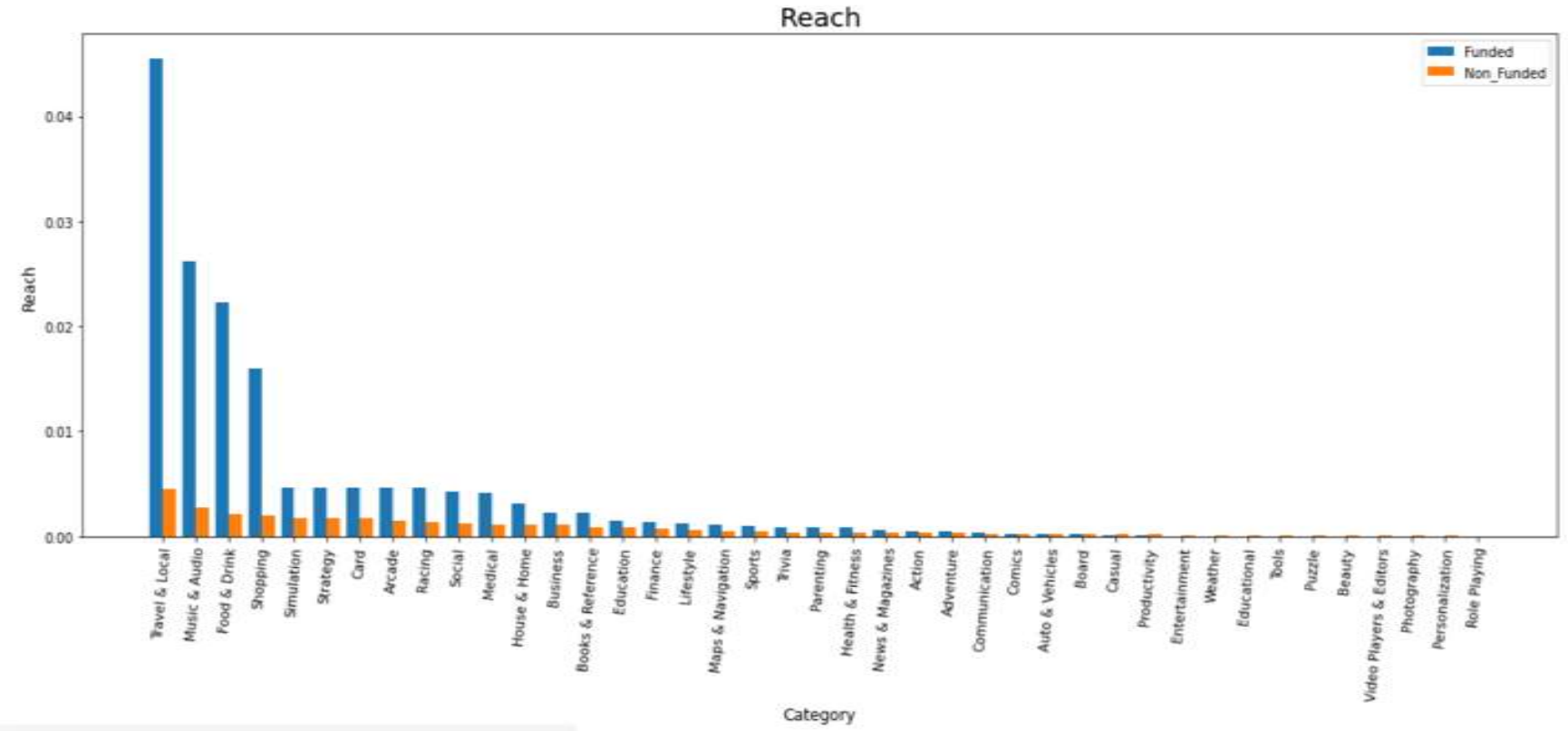
Exploratory Data Analysis



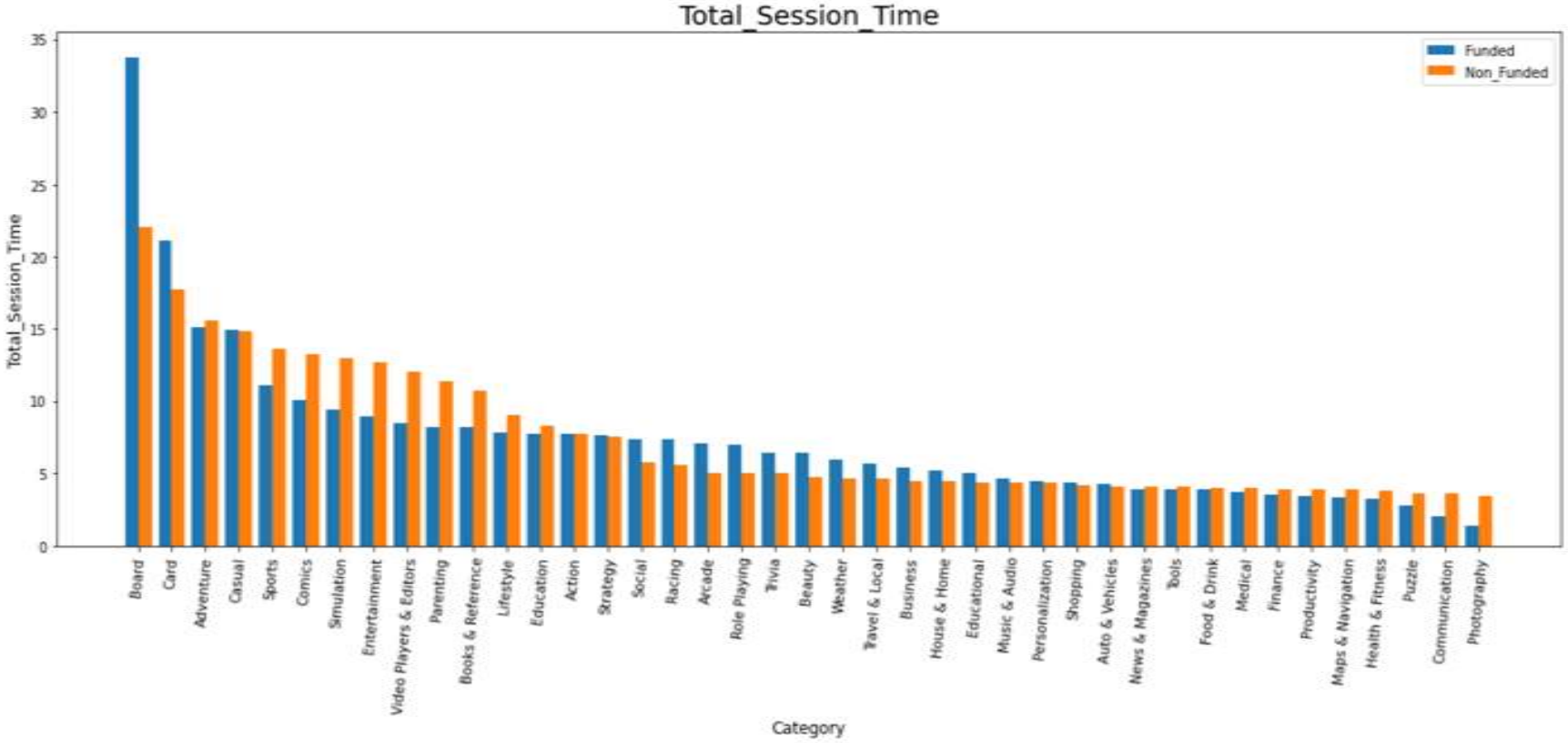
Exploratory Data Analysis



Exploratory Data Analysis



Exploratory Data Analysis



Feature Engineering

Avg_Session_Time_Prev3 : The average session time upto 3rd last month.

Open_Rate_Prev3 : The open rate upto 3rd last month.

Reach_Prev3 : The reach upto 3rd last month.

Total_Session_Time_Prev3 : The total session time upto 3rd last month.

Feature Engineering(Contd.)

Avg_Session_Time_change_3 : The expected change in average session time after 3 months.

Avg_Session_Time_change_2 : The expected change in average session time after 2 months.

Avg_Session_Time_change_1 : The expected change in average session time after 1 month.

Open_Rate_change_3 :The expected change in open rate after 3 months.

Open_Rate_change_2 : The expected change in open rate after 2 months.

Feature Engineering(Contd.)

Open_Rate_change_1 : The expected change in open rate after 1 month.

Reach_change_3 : The expected change in Reach after 3 months.

Reach_change_2 : The expected change in Reach after 2 months.

Reach_change_1 : The expected change in Reach after 1 month.

Total_Session_Time_change_3 : The expected change in total session time after 3 months.

Feature Engineering(Contd.)

Total_Session_Time_change_2 : The expected change in total session time after 2 months.

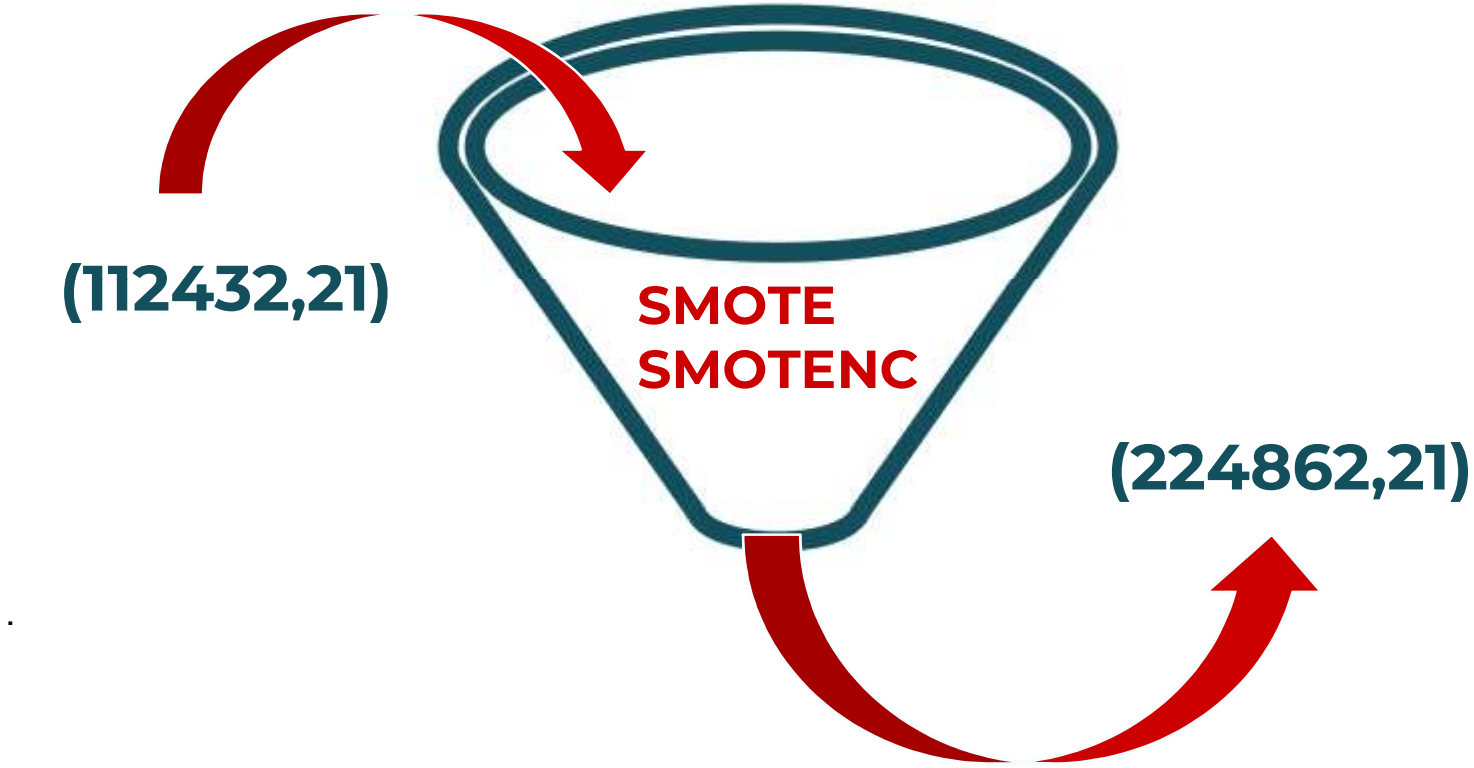
Total_Session_Time_change_1 : The expected change in total session time after 1 month.

Times_Funded_prev : Number of time the startup got funded earlier.

Investor_Interest : The percentage of interest ,the investor shows.

Funded : Dependent variable indicating whether startup will get funded in next 3 months.

Data Pre-processing for Model



Applying ML Algorithms

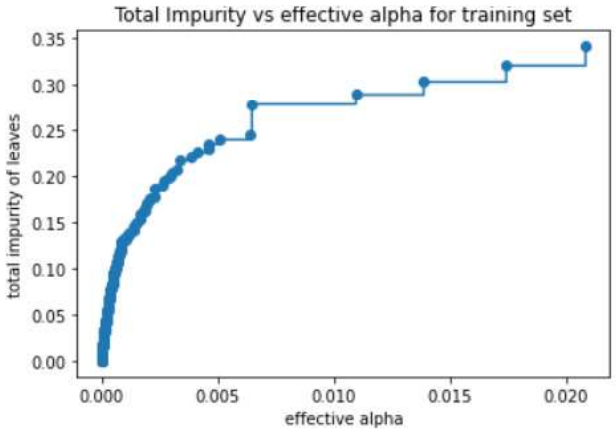
Stochastic Gradient Descent

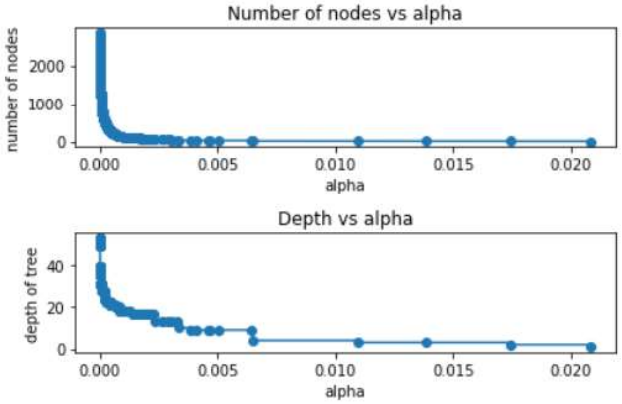
```
Training accuracy Score    :  0.9984292843775064
Validation accuracy Score :  0.9983067462792977
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     22270
           1       1.00      1.00      1.00     22614

    accuracy                           1.00     44884
   macro avg       1.00      1.00      1.00     44884
weighted avg       1.00      1.00      1.00     44884
```

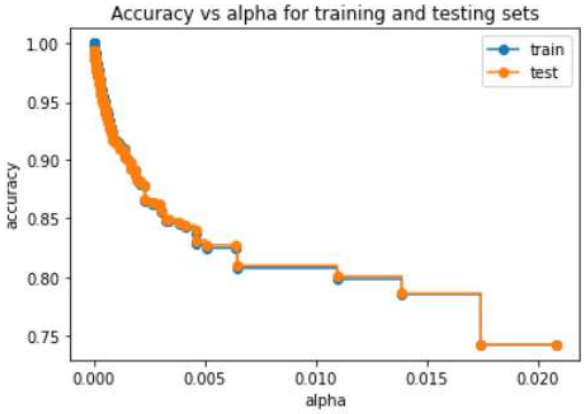
model	train_accuracy	train_precision	train_recall	train_f1_score	train_tn fp fn tp	train_auc_roc	test_accuracy	test_precision	test_recall	test_f1_score	test_tn fp fn tp	test_auc_roc
SGD	0.998429	0.996865	1.0	0.99843	[89582, 282, 0, 89672]	0.998431	0.998307	0.996639	1.0	0.998317	[22270, 76, 0, 22538]	0.998299

Applying ML Algorithms

Decision Tree Classifier(Cost Complexity Pruning)







Applying ML Algorithms

Performance of Decision Tree Classifier

model	train_accuracy	train_precision	train_recall	train_f1_score	train_tn fp fn tp	train_auc_roc	test_accuracy	test_precision	test_recall	test_f1_score	test_tn fp fn tp	test_auc_roc
Decision Tree	0.999983	1.0	0.999967	0.999983	[89864, 0, 3, 89669]	0.999983	0.993873	0.990742	0.997116	0.993919	[22136, 210, 65, 22473]	0.993859

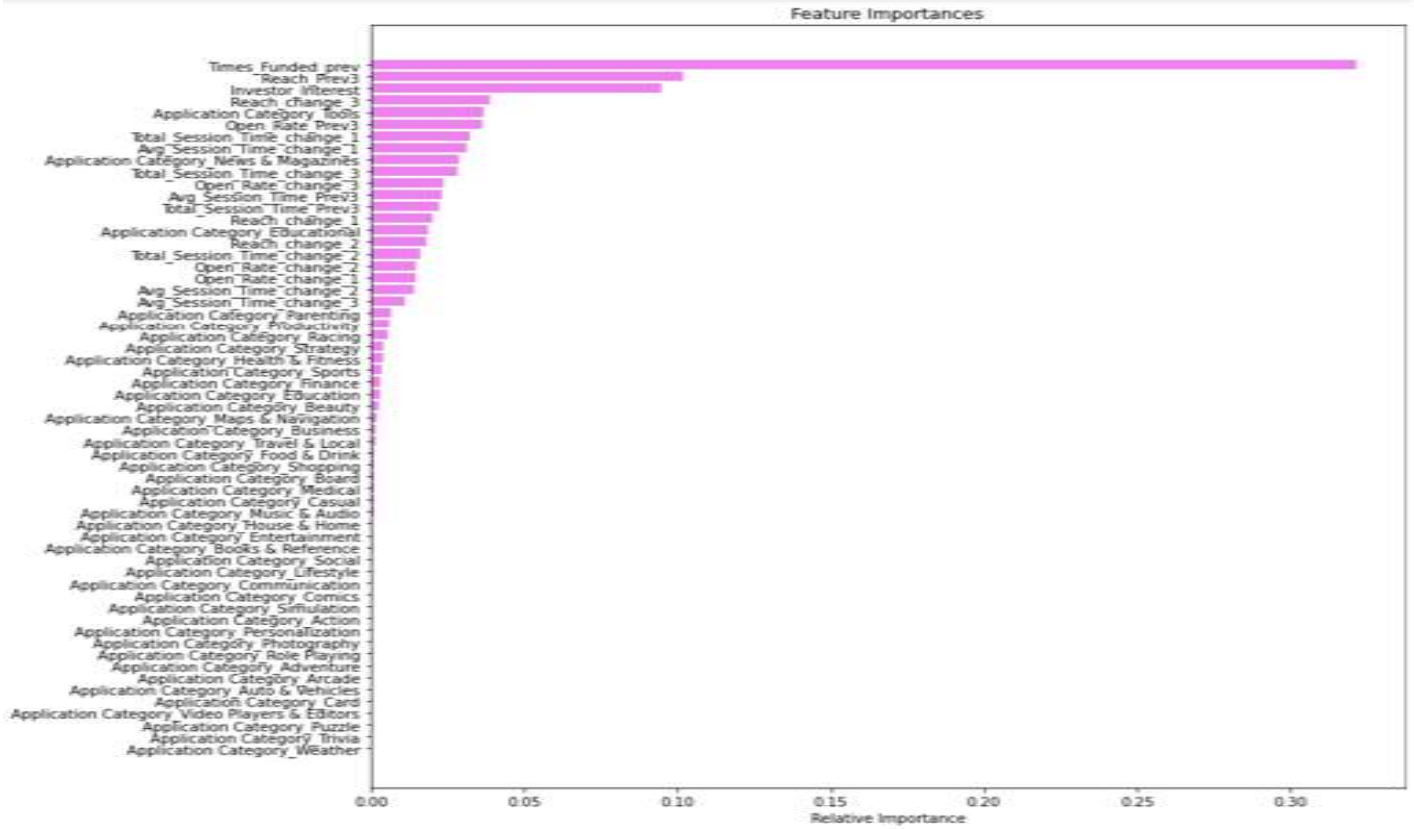
Performance of Gradient Boosting Classifier

model	train_accuracy	train_precision	train_recall	train_f1_score	train_tn fp fn tp	train_auc_roc	test_accuracy	test_precision	test_recall	test_f1_score	test_tn fp fn tp	test_auc_roc
GBM	0.947983	0.96379	0.930826	0.947021	[86728, 3136, 6203, 83469]	0.947964	0.944925	0.963418	0.925459	0.944057	[21554, 792, 1680, 20858]	0.945008

Performance of XG Boost Classifier

model	train_accuracy	train_precision	train_recall	train_f1_score	train_tn fp fn tp	train_auc_roc	test_accuracy	test_precision	test_recall	test_f1_score	test_tn fp fn tp	test_auc_roc
XGB	0.94156	0.959718	0.921681	0.940315	[86395, 3469, 7023, 82649]	0.941539	0.94214	0.960552	0.922664	0.941227	[21492, 854, 1743, 20795]	0.942223

Feature Importance



Cat Boost

Training Classification Report

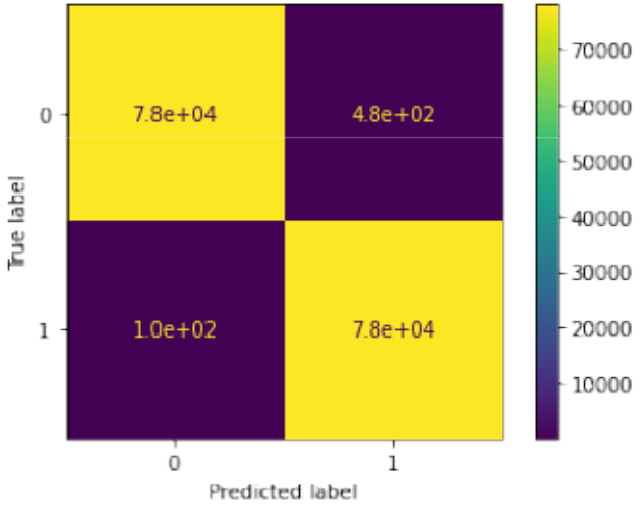
	precision	recall	f1-score	support
0	1.00	0.99	1.00	78727
1	0.99	1.00	1.00	78367
accuracy			1.00	157094
macro avg	1.00	1.00	1.00	157094
weighted avg	1.00	1.00	1.00	157094

Test Classification Report

	precision	recall	f1-score	support
0	1.00	0.99	1.00	33483
1	0.99	1.00	1.00	33843
accuracy			1.00	67326
macro avg	1.00	1.00	1.00	67326
weighted avg	1.00	1.00	1.00	67326

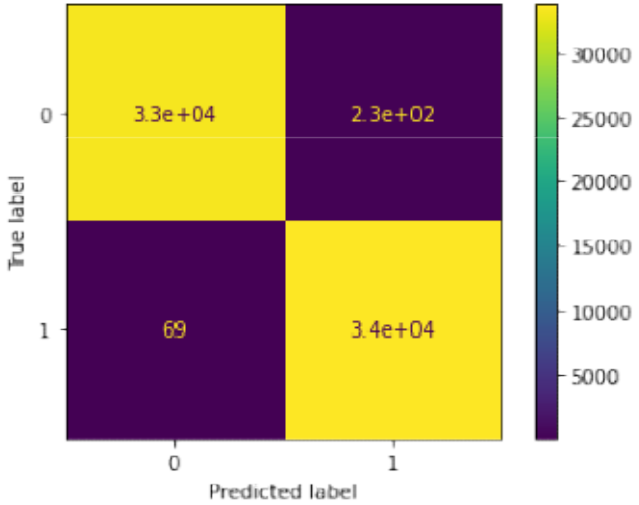
Cat Boost

Confusion Matrix for Training data



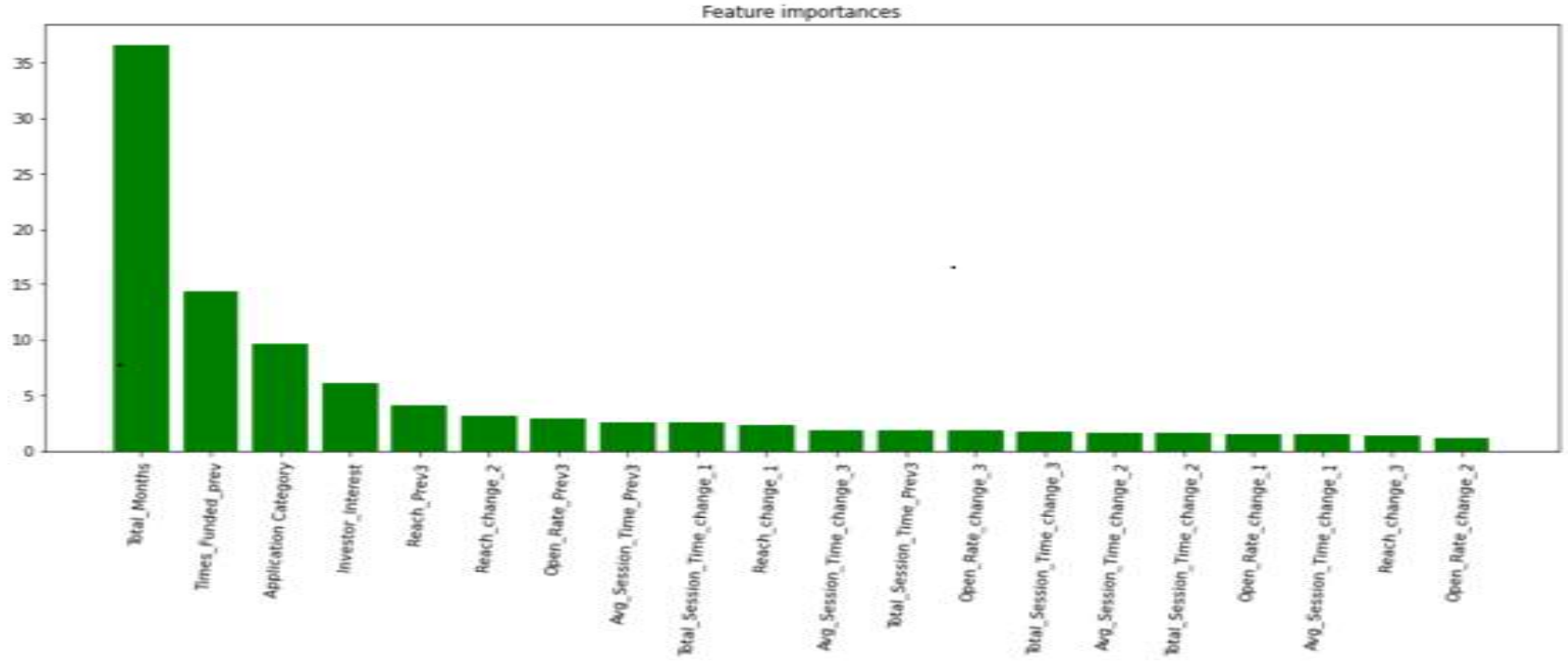

```
[78242,   485],
       [  105, 78262]
```

Confusion Matrix for Test data

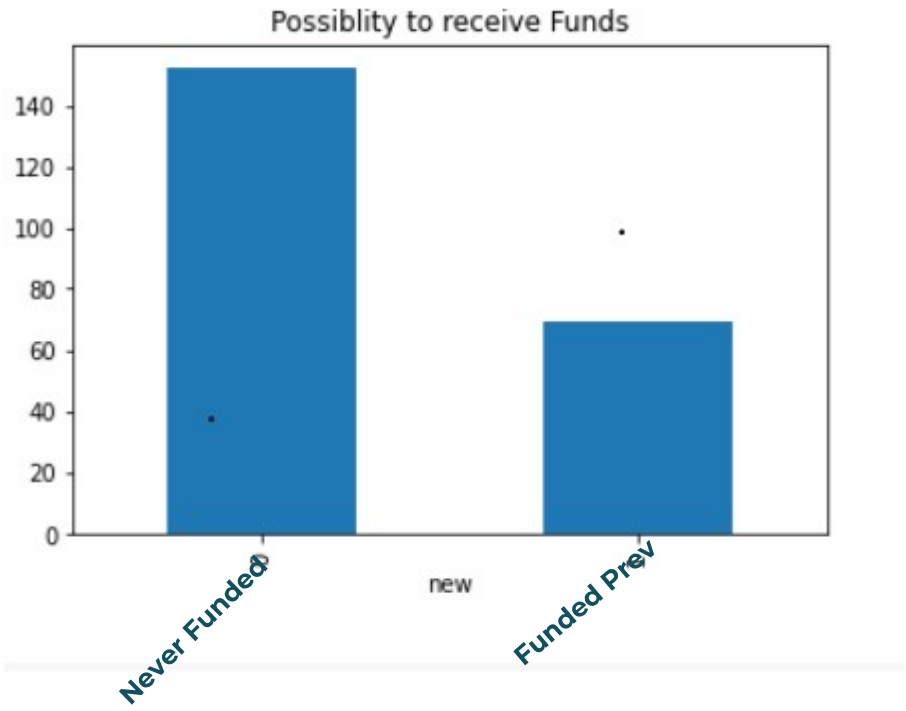



```
[33252,   231],
       [   69, 33774]
```

Cat Boost


Feature importances
35
30
25
20
15
10
5
0
Total_Months
Times_Funded_prev
Application Category
Investor_Interest
Reach_Prev3
Reach_change_2
Open_Rate_Prev3
Avg_Session_Time_Prev3
Total_Session_Time_change_1
Reach_change_1
Avg_Session_Time_change_3
Total_Session_Time_Prev3
Open_Rate_change_3
Total_Session_Time_change_3
Avg_Session_Time_change_2
Total_Session_Time_change_2
Open_Rate_change_1
Avg_Session_Time_change_1
Reach_change_3
Open_Rate_change_2

Revisiting the Problem



- Periodic Investment
- Apply with other Domain/ Category
- Age of company
- Relations with investors
- Considering other features while investing more than once.

Never Funded (Cat Boost)

Training Classification Report

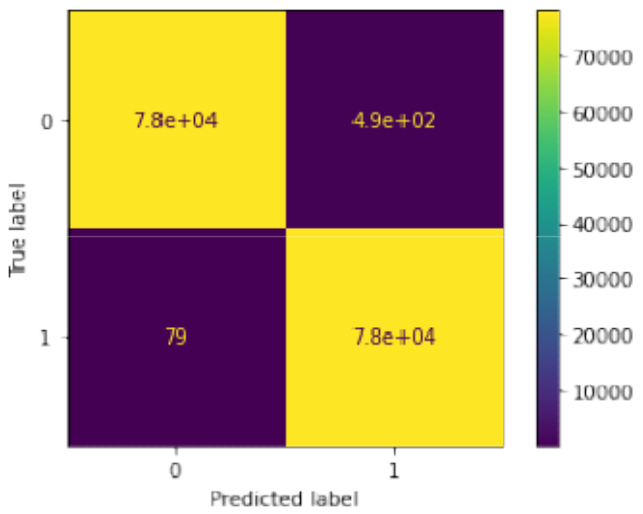
	precision	recall	f1-score	support
0	1.00	0.99	1.00	78412
1	0.99	1.00	1.00	78180
accuracy			1.00	156592
macro avg	1.00	1.00	1.00	156592
weighted avg	1.00	1.00	1.00	156592

Test Classification Report

	precision	recall	f1-score	support
0	1.00	0.99	1.00	33440
1	0.99	1.00	1.00	33672
accuracy			1.00	67112
macro avg	1.00	1.00	1.00	67112
weighted avg	1.00	1.00	1.00	67112

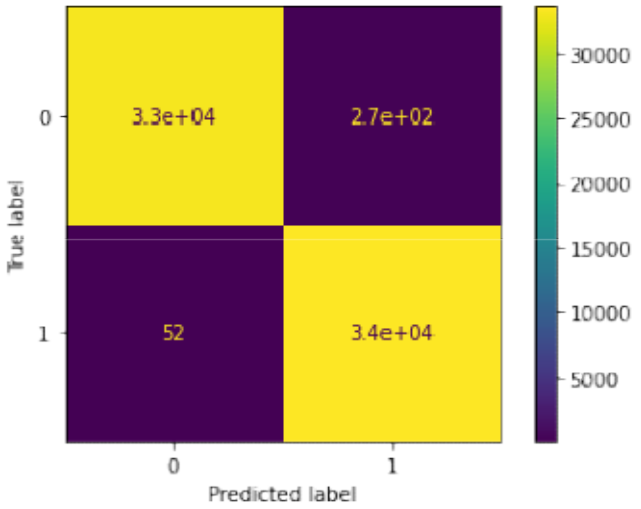
Never Funded (Cat Boost)

Confusion Matrix for Training data



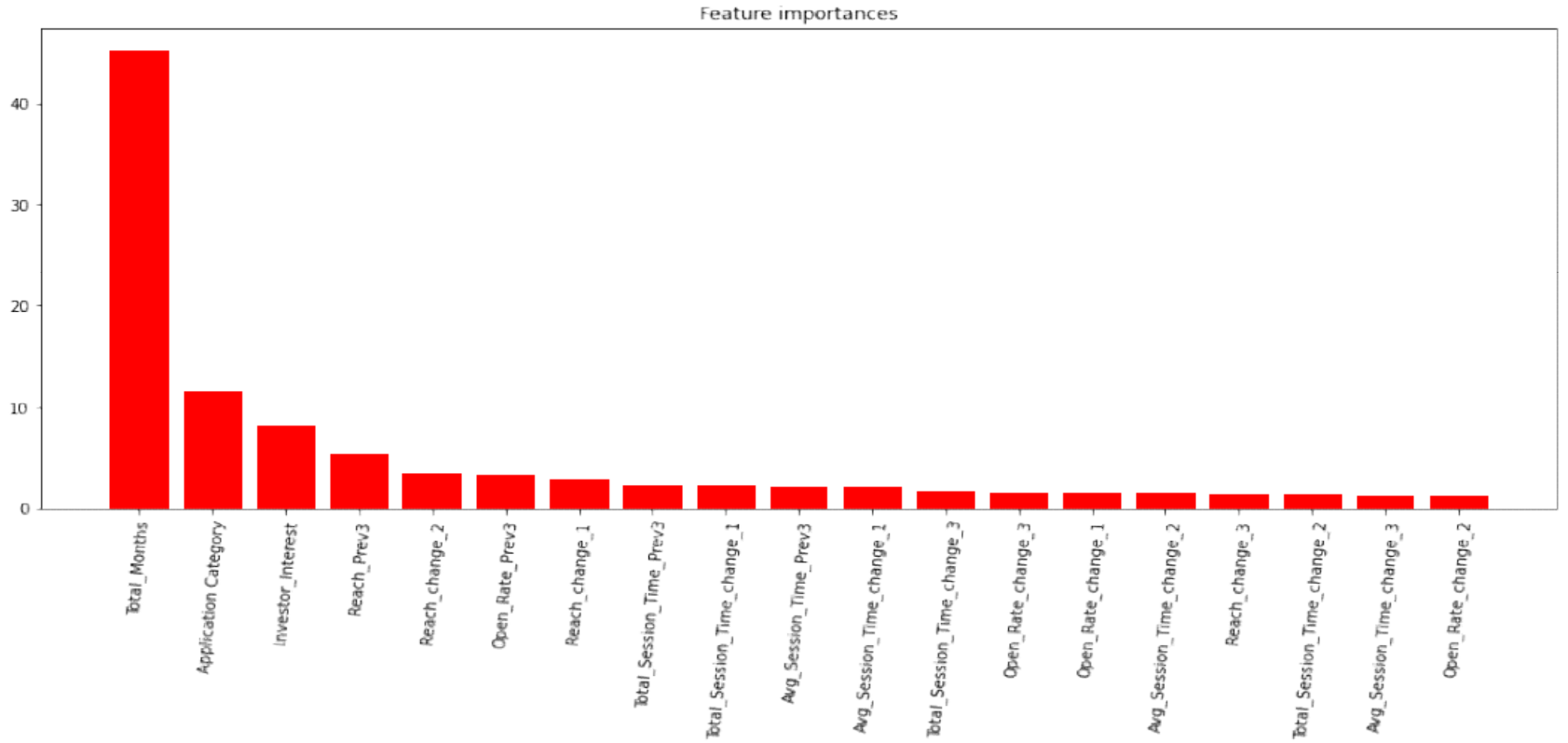

```
[77921,   491],
       [   79, 78101]
```

Confusion Matrix for Test data




```
[33170,   270],
       [   52, 33620]
```

Feature Importance



Funded Previously (Cat Boost)

Training Classification Report

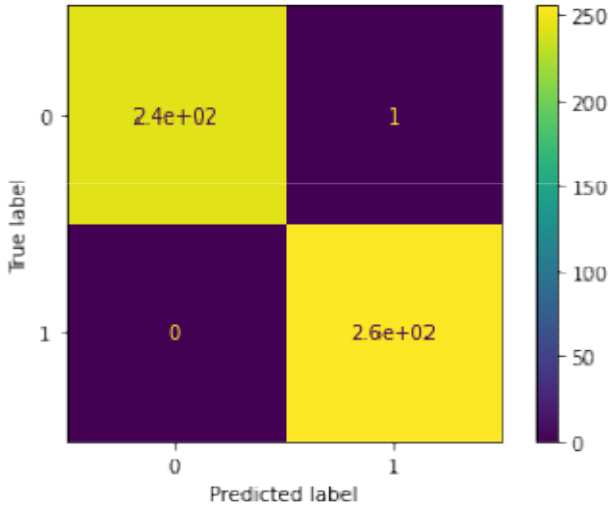
	precision	recall	f1-score	support
0	1.00	1.00	1.00	245
1	1.00	1.00	1.00	256
accuracy			1.00	501
macro avg	1.00	1.00	1.00	501
weighted avg	1.00	1.00	1.00	501

Test Classification Report

	precision	recall	f1-score	support
0	0.96	0.96	0.96	113
1	0.96	0.96	0.96	102
accuracy			0.96	215
macro avg	0.96	0.96	0.96	215
weighted avg	0.96	0.96	0.96	215

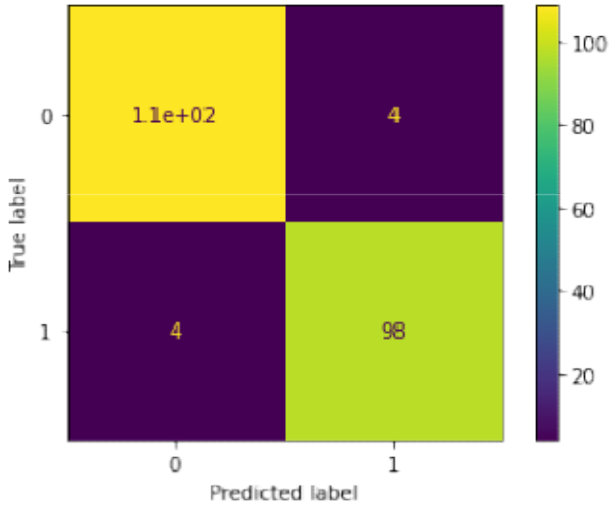
Funded Previously (Cat Boost)

Confusion Matrix for Training data



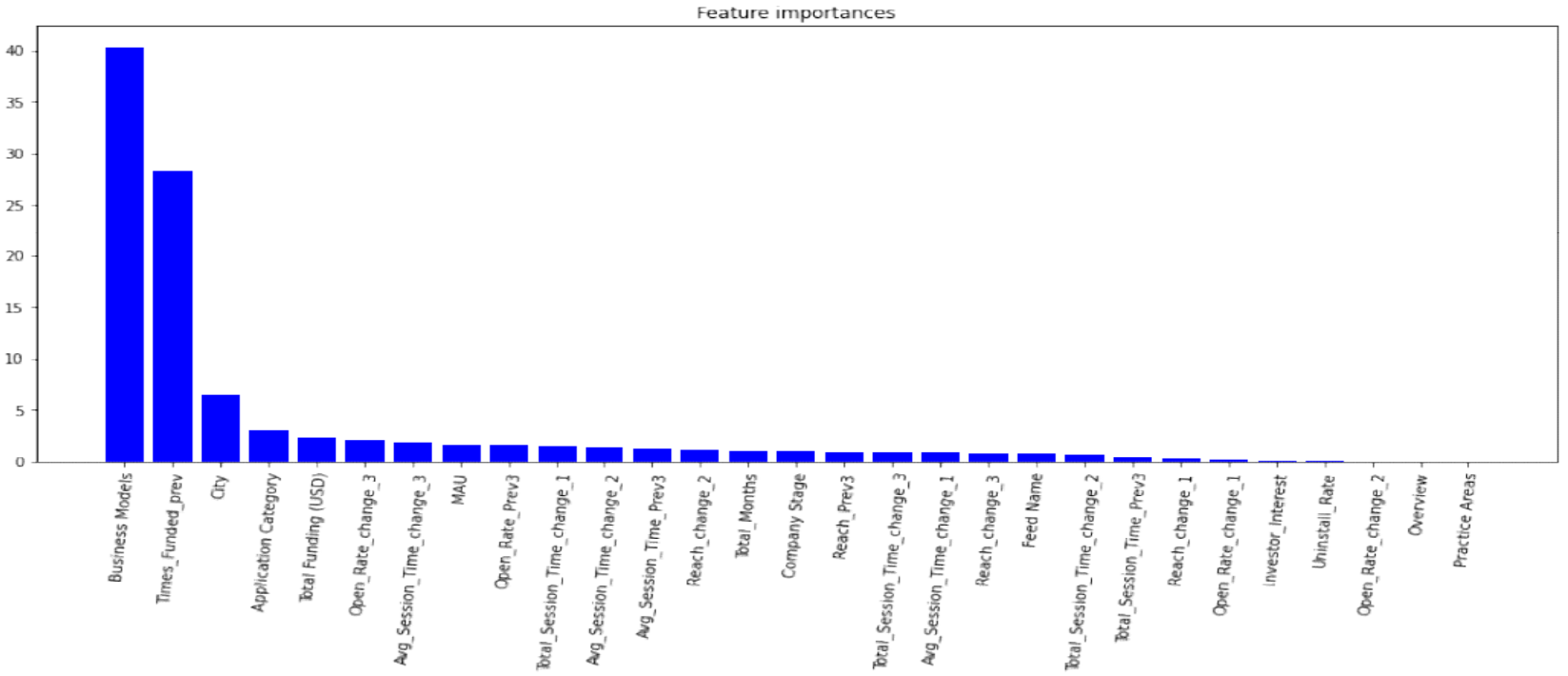

```
[244,   1],
[  0, 256]
```

Confusion Matrix for Test data




```
[109,   4],
[  4,  98]
```

Feature Importance



Conclusion

- As the Exploratory Data Analysis suggests,maximum seeding funding is given to startups.
- In the year, 2020 total amount of funding was maximum.
- The Category which got funding for maximum number of times was Education.
- Maximum funding was given to Bangalore based startups.
- As we had other more feature for startups that were funded once, for this problem we suggest to build two separate ML systems. (Even can do multiclass classification)
- Cat Boost was performing well overall. So optimal model would be same for both ML systems.

Challenges

- The dataset was too large to handle.
- There were plenty of missing values.
- We had to build relevant features to approach the problem statement.
- As the dataset was very large, it took great deal of time to prepare the training dataset.
- Due to lack of time we were unable to make good model for startups that were already funded.

Q & A