Capstone Project - 4

Startup Funding Prediction

Suraj Pandey

The Entrepreneurship journey

1. Problem Statements
2. Data Summary
3. Analysis of Data
4. Null values Imputation/ Data Cleaning
5. Data Preparation
6. Feature Engineering
7. Model Training
8. Evaluation Metrics
9. Challenges
10. Conclusion

Problem Statement

1. 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.

1. The objective of the project is to come up with a machine learning model to predict whether a startup will get funded in the next three months using app traction data and startup details.

Data Summary

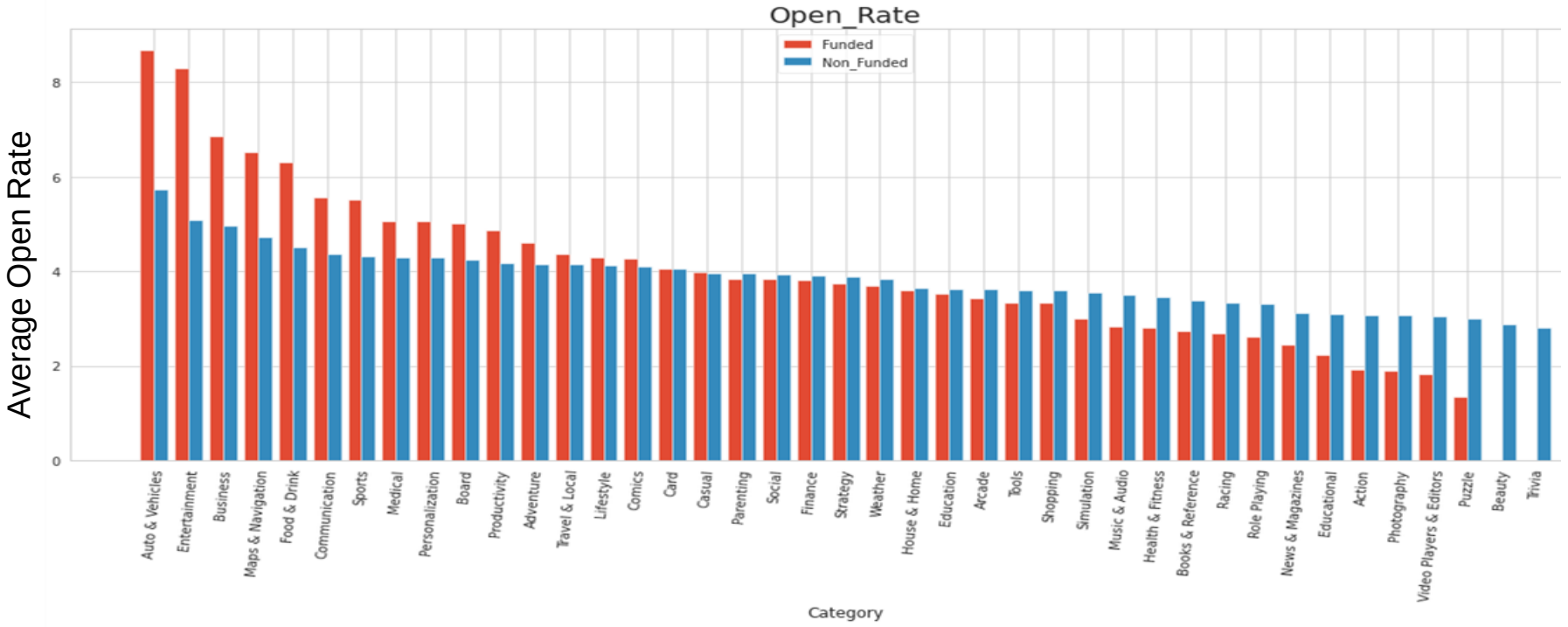
The dataset spans over three years from July 2018 to August 2020.

The dataset contains 20 features with more than 15 lakh observations.

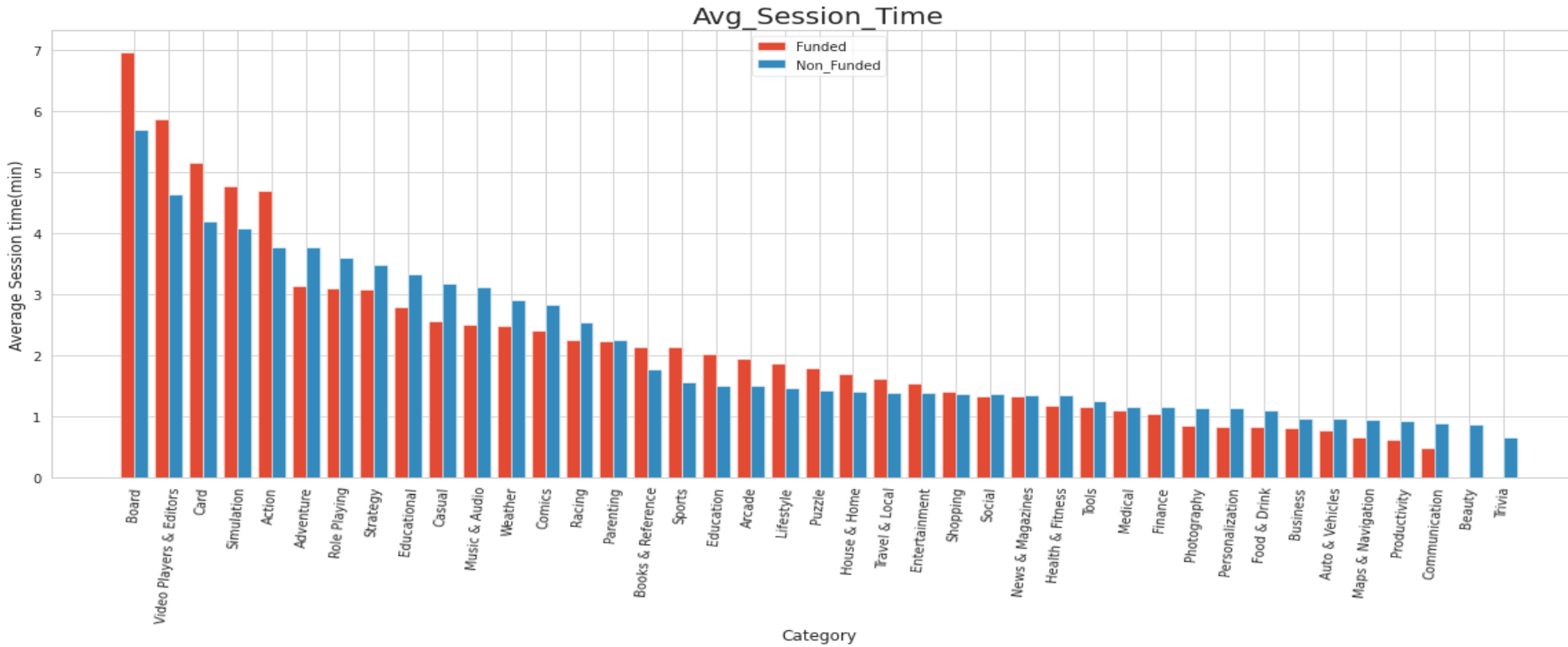
Few Features From Dataset

1. UUID :- Unique identifier for a single startup entity.
2. Month :- Month for which the app data is available.
3. Application Category :- The category to which an application belongs to
4. Average Session Time :- Average time of the session in app during the month (in minutes)
5. Total Session Time :- Average Session Time/user * open rate
6. Open Rate :- No. of times app has been opened by a user.
7. Reach :- % of devices having the app installed
8. Funding_ind :- Indicator for a funded startup

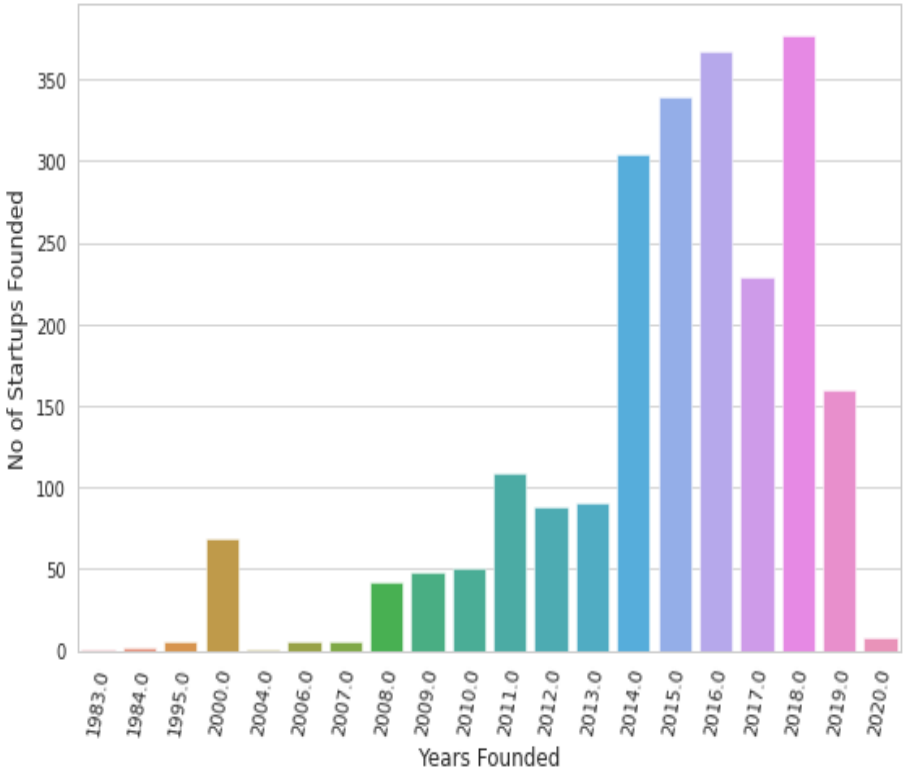
Mean Open Rate (By Application Category)



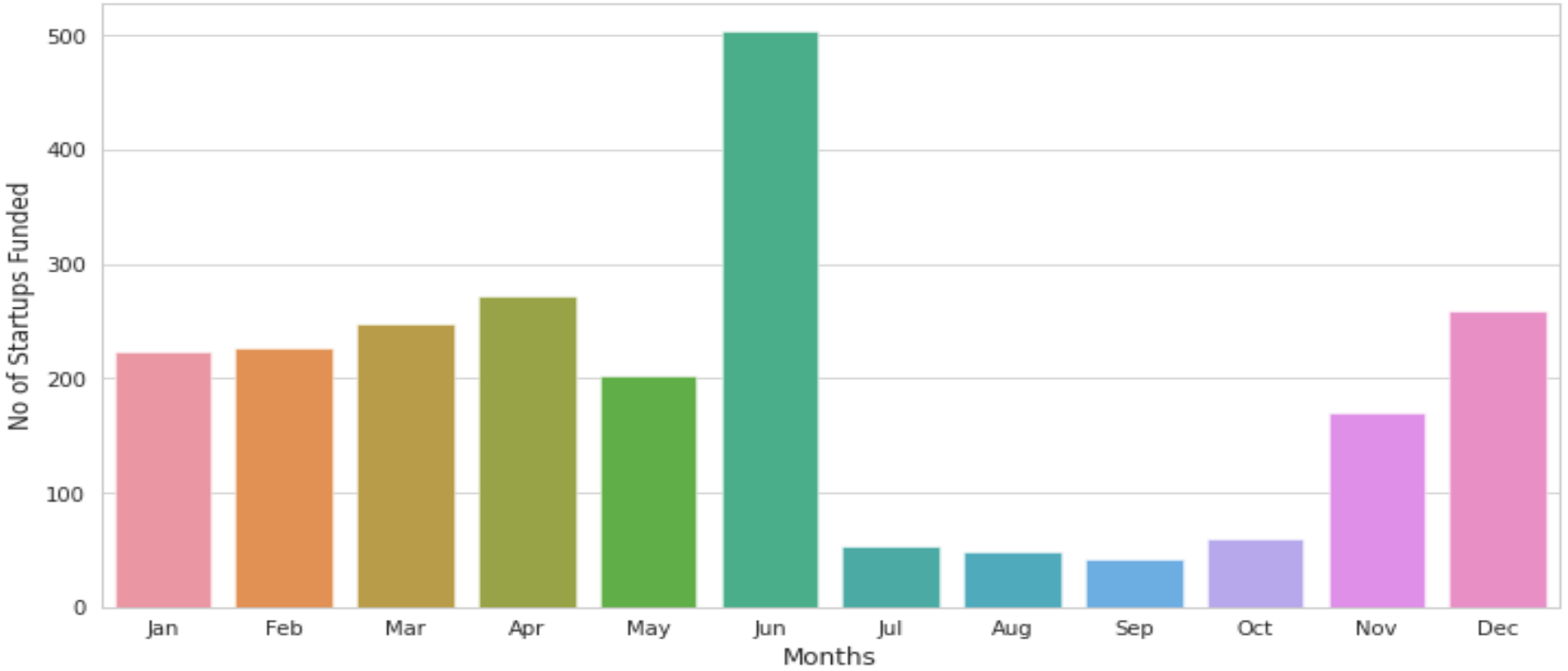
Mean Session Time (By Application Category)



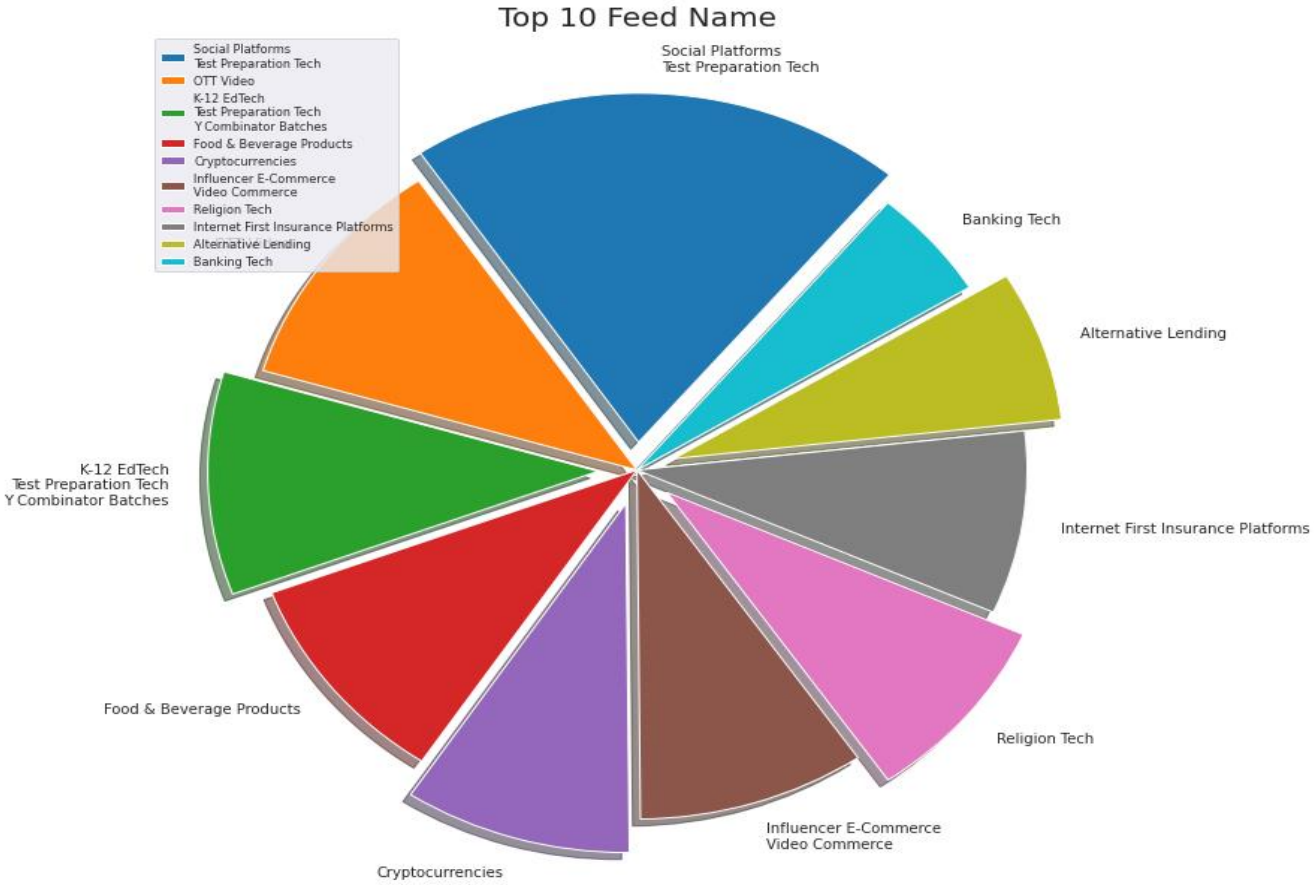
Growth in Startup Ecosystem



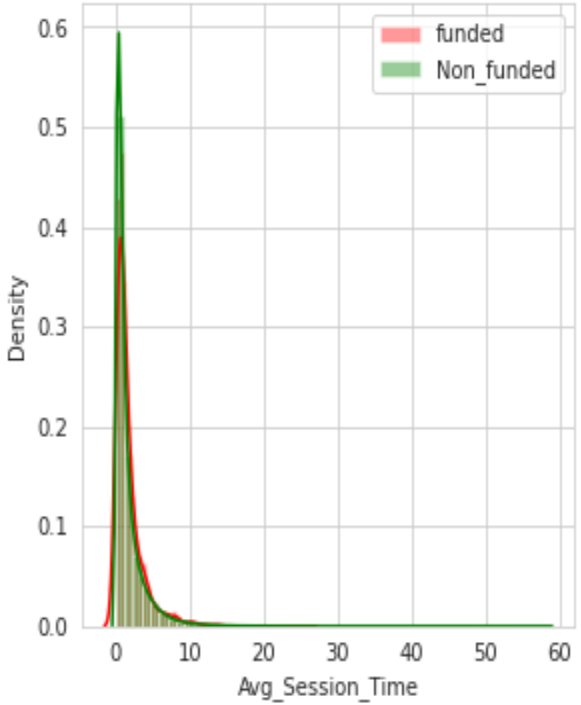
Which Month Brings More Funds?

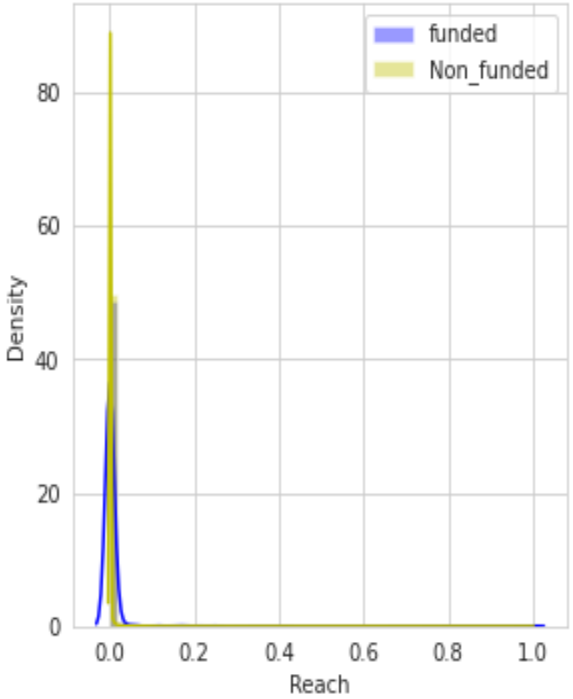


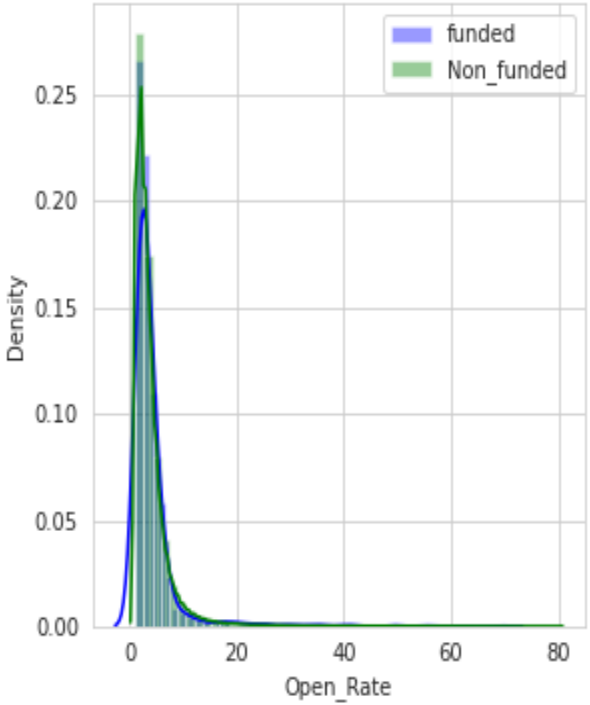
Top 10 Funded Business Domains



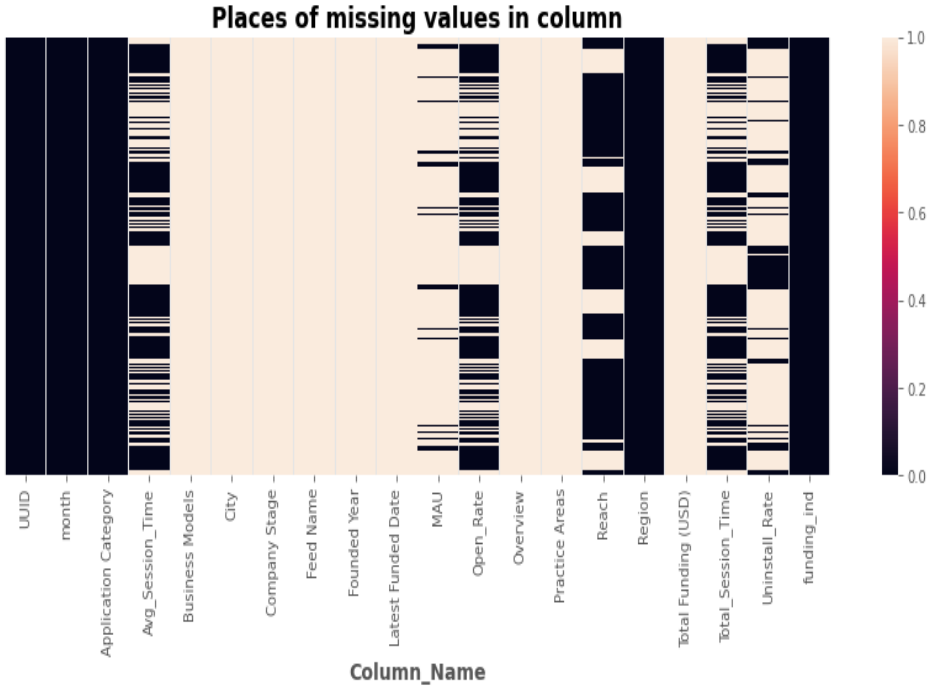
Distribution of app traction features







Where are the missing values ?



Data Cleaning

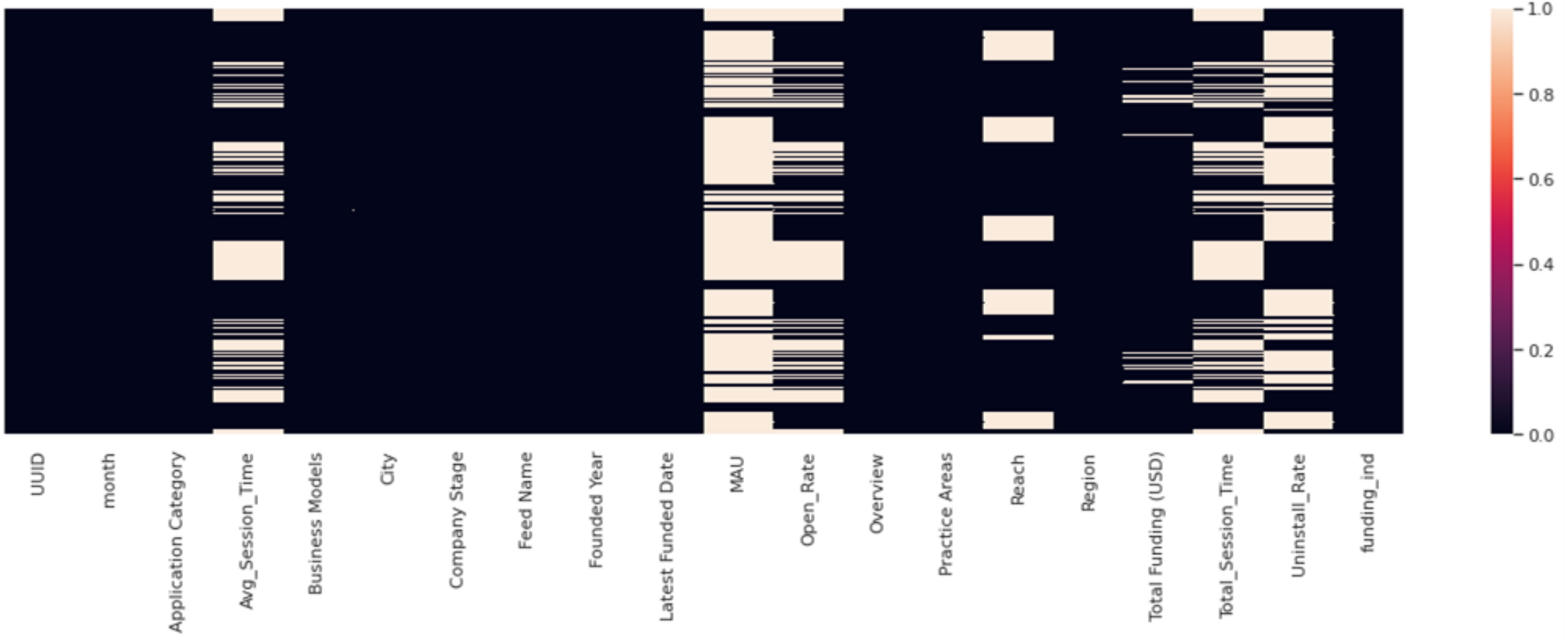
1. Most of the features contain more than 90% null values.
2. App traction features are selected for data cleaning includes Average Session Time, Total Session Time, Open Rate, Reach, funding indicator.
3. Interpolation for imputation followed by forward fill and backward fill.

Interpolation

1	na	na	4	4

1	2	3	4	4

Missing values in funding_ind = 1



Data Preparation

1. The features we have now aren't enough to predict.
2. We created features having the data for previous three months.
3. There are multiple observations for particular month.
4. We took the mean of the data grouping by UUID and month.
5. We Created previous month features for each observation.
6. We discarded the rows which do not have previous three months of data.

Data Preparation(contd.)

While Grouping by month and UUID, we will take the mean of Avg_Ses_Time

UUID	month	Avg_Ses_Time
1	Jan	5.5
1	Jan	4.5
1	Feb	4.5
2	Jun	5.0
2	Jun	7.0
3	Jul	7.0

Data Preparation(contd.)

After taking the mean for each month and UUID pair, we will be left with these.

***If the mean of target is greater than zero, then it's taken to be 1**

UUID	month	Avg_Ses_Time
1	Jan	5.0
1	Feb	4.5
2	Jun	6.0
3	Jul	7.0

Rows before the operation	Rows after the operation
1502175	757290

Data Preparation(contd.)

Rotating the columns UUID and Avg_Ses_Time

UUID	UUID_1	month	Avg_Ses_Time	Avg_Ses_Time_Prev
1	3	Jan	5.0	7.0
1	1	Feb	4.5	5.0
2	1	Jun	6.0	4.5
3	2	Jul	7.0	6.0

Data Preparation(contd.)

- We are now left with rows which have previous month of data and the previous month's averaged characteristics are now appended as new features

UUID	UUID_1	month	Avg_Ses_Time	Avg_Ses_Time_Prev
1	1	Feb	4.5	5.0

Rows before the operation	Rows after the operation
757290	549319

Feature Engineering

1) Temporally Decayed Weighted Averages of app traction features

1. Took a look back window of 3 months.
1. Weights decayed based on time.
1. Weighted average by given below equation.

Time Decayed Weighted Average=(month1*0.5)+(month2*0.33)+(month3*0.16)

2) Category Dominance

Category dominance features tell us how a particular company performs with respect to the competitors.

Dominance = (Average session time of company X) / (Total Average session time of all the companies in the category of company X)

3) Instability Metric

Instability metric captures the fluctuations in the app traction metrics of a company.

Residuals = (Values of feature vector) - (Mean of the feature Vector)

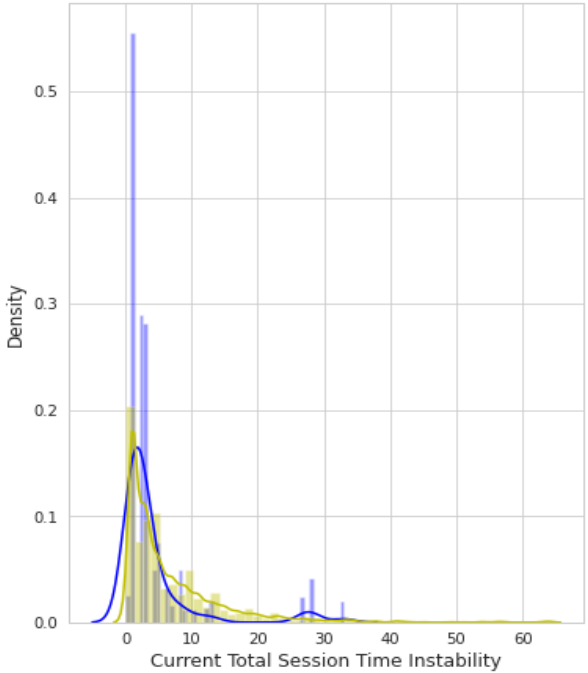
Instability Metric = No. of time sign changes successively

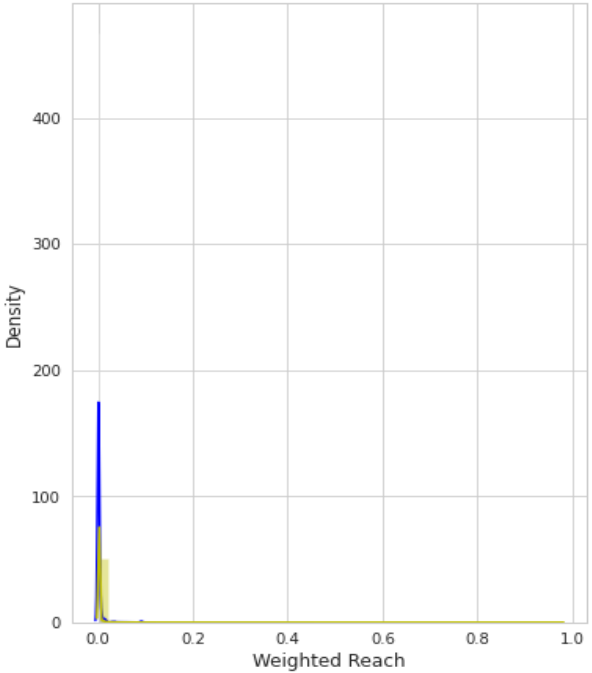
4) MoM Percentage Change in Traction

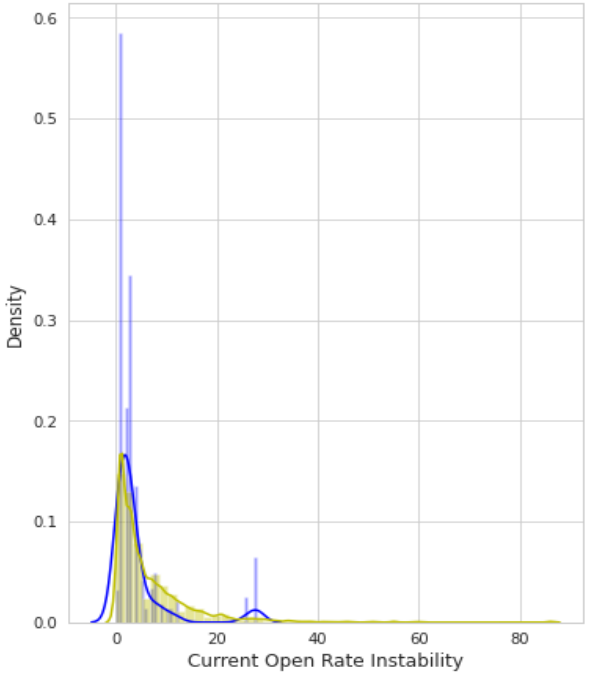
Percentage change in app traction features with respect to previous month's records captures the growth of a company.

% change in Reach = ((Current month Reach) - (Previous month Reach))/ (Previous month Reach)

New Features Distribution





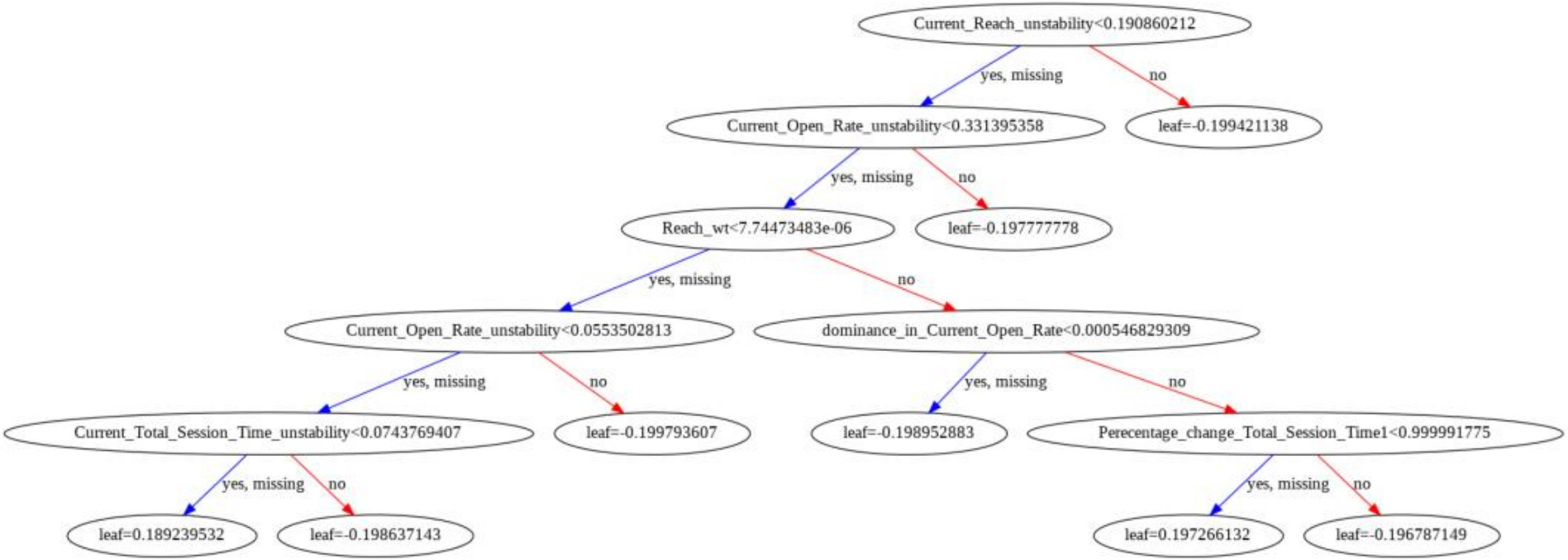


Definitions of top predictors

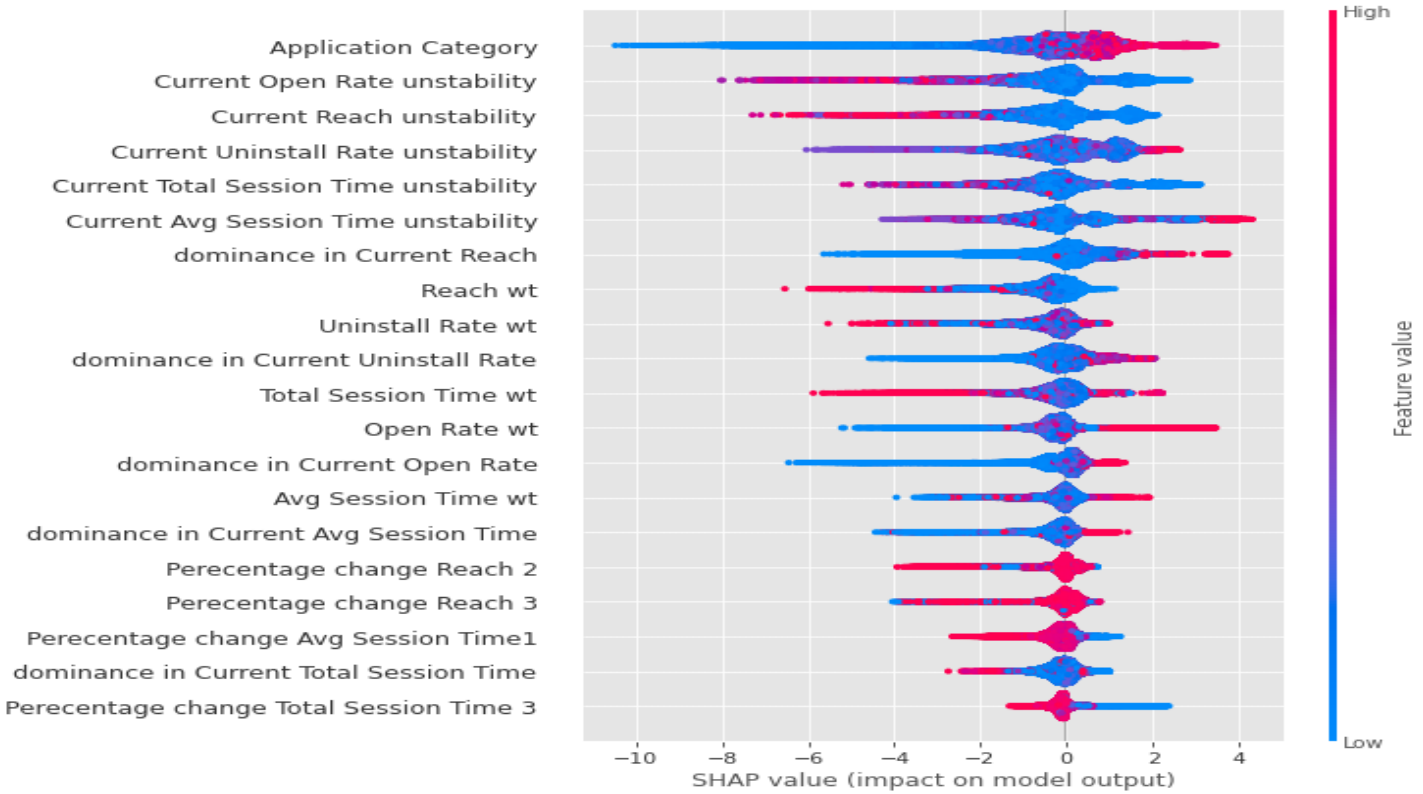
1.	Current Open Rate instability	Instability in trend for company's open rate.
2.	Current reach instability	Instability in trend for company's reach
3.	Current total session Time instability	Instability in trend for company's Total Session Time
4.	Weighted Average Reach	We have given more weightage to recent months as compared to previous months and then calculated the average of the reach based on all the previous records.

5.	Dominance in Current Total Session Time	Indicates the dominance level of a company in its application category based on it's Total Session Time
6.	Dominance in Current Reach	Indicates the dominance level of a company in its application category based on it's Reach
7.	Dominance in Current Open rate	Indicates the dominance level of a company in its application category based on it's Open rate
8.	% change in open_rate	Percentage change in open rate with respect to previous months open rate
9.	% change in Uninstall Rate	Percentage change in Uninstall with respect to previous months Uninstall rate

Simple Decision Tree



Impact of key drivers on model predictions



Objective : High Recall Predictive Model

As a venture capitalist, people seek to invest in startups with high ROI, so we are trying to make a model that should accurately predict startups that can get funding in the upcoming three months.

That is to have high recall

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

We wanted to have high Recall value for Funded Startups. We gave 90% weightage for correct classification of 1 i.e Funded Startups.

Challenges

1. Huge dataset with lots of missing and redundant values.

1. Filtering was tricky.

1. Imputing null values was very challenging and time taking process.

1. ML problem formulation was tricky.

Conclusion

1. Overall accuracy is 98%.

1. With 68% of recall, we can predict the start-ups that will receive fundings in the next three months.

Q & A