Kickstarter Campaigns Analysis

Predicting Success with Machine Learning

CMPE 257

Prof Dr. Jahan Ghofraniha

San Jose State University

Team Members

Darshan Jani

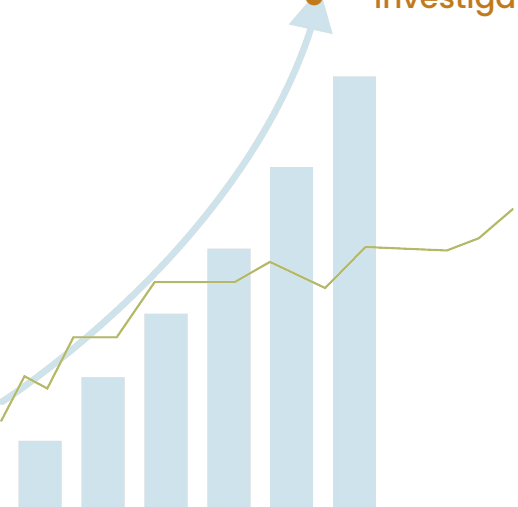
Rohit Sharma

Charika Bansal

Problem Statement

- Analyzing Kickstarter campaigns dataset to predict the success of a campaign using only information from project launch

- Investigating the relationship between funding and GDP (Add-on)



Motivation

- Importance of understanding factors contributing to successful crowdfunding campaigns
- Analyzing GDP impact on funding to better understand the economic landscape
- Improving prediction models for future campaigns

Dataset

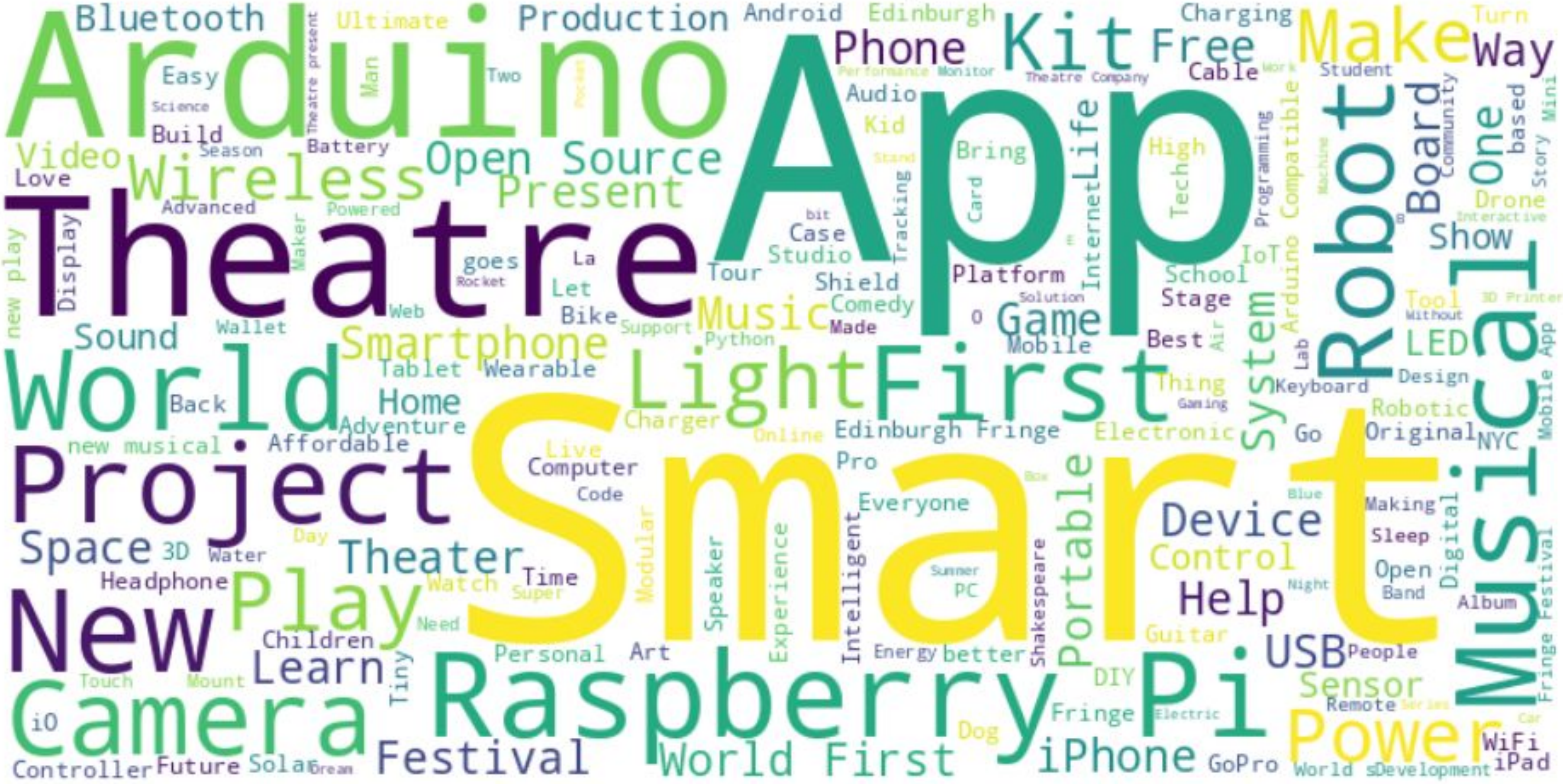
- Kickstarter campaigns dataset as of February 1st, 2017, containing 20,632 records with attributes such as funding goal, project name, blurb, pledged amount, backers, state, deadlines, and more

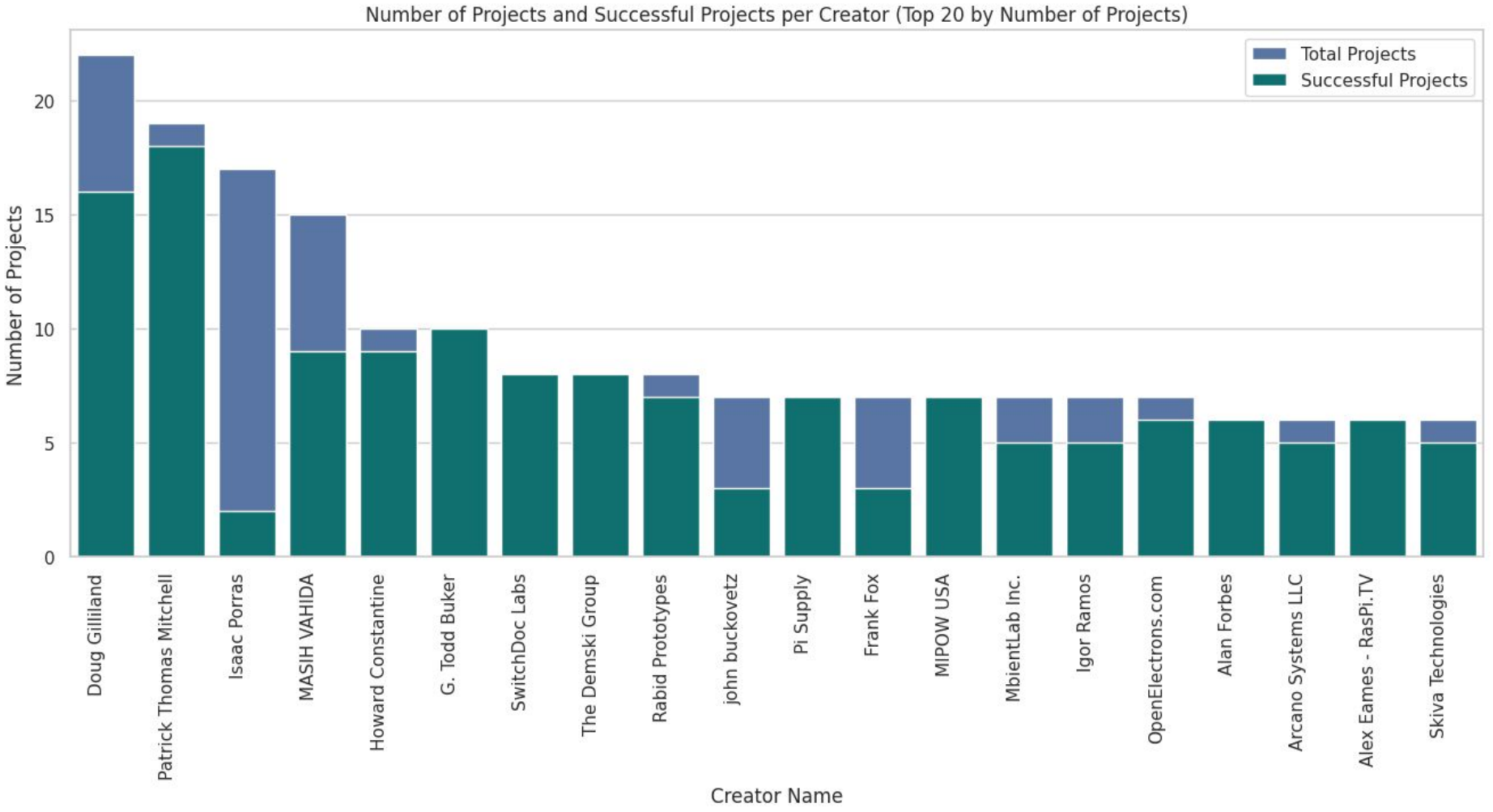
- Source: Kaggle (https://www.kaggle.com/datasets/sripaadsrinivasan/kickstarter-campaigns-dataset)

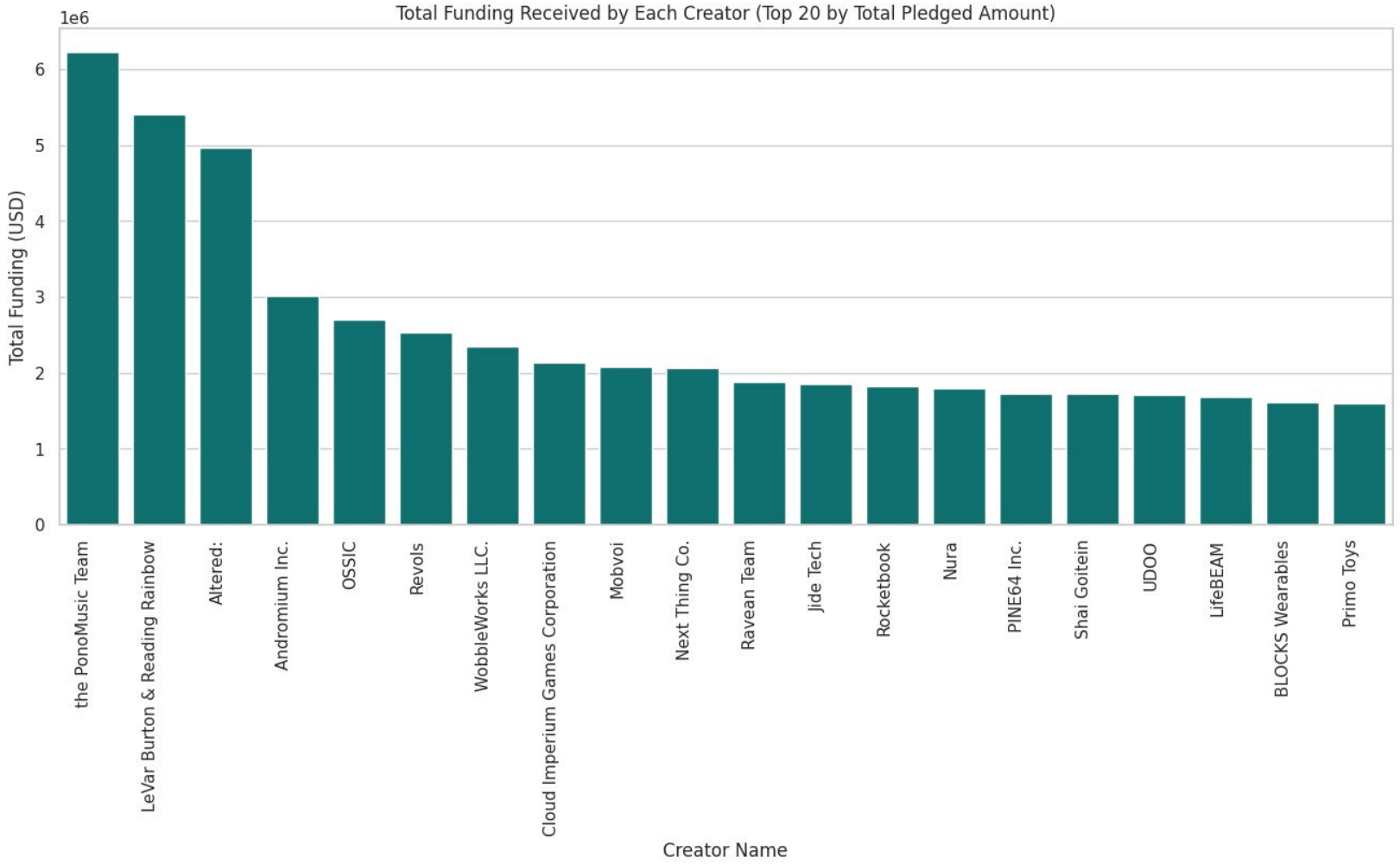
- Additional data: GDP for respective countries and years

- Source: World Bank (https://data.worldbank.org/indicator/NY.GDP.MKTP.CD)

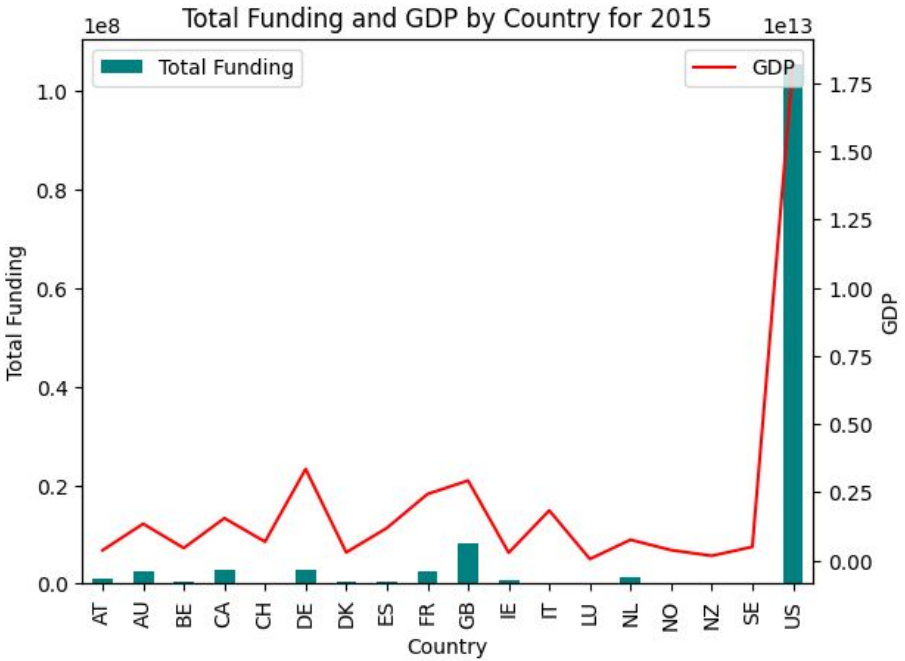
Word Cloud of Successful Project Names



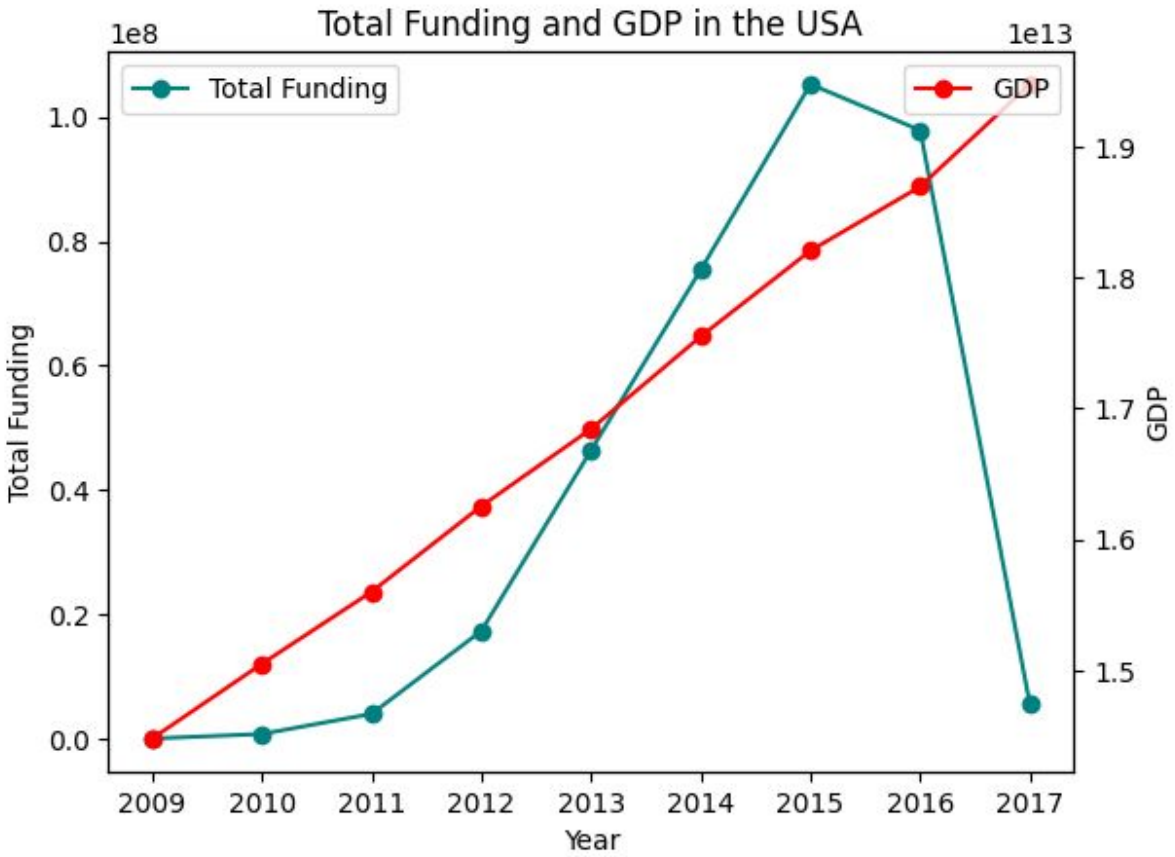

Number of Projects and Successful Projects per Creator (Top 20 by Number of Projects)
Total Projects
Successful Projects
Number of Projects
0
5
10
15
20
Doug Gilliland
Patrick Thomas Mitchell
Isaac Porras
MASIH VAHIDA
Howard Constantine
G. Todd Buker
SwitchDoc Labs
The Demski Group
Rabid Prototypes
john buckovetz
Pi Supply
Frank Fox
MIPOW USA
MbientLab Inc.
Igor Ramos
OpenElectrons.com
Alan Forbes
Arcano Systems LLC
Alex Eames - RasPi.TV
Skiva Technologies
Creator Name


Total Funding Received by Each Creator (Top 20 by Total Pledged Amount)
1e6
Total Funding (USD)
0
1
2
3
4
5
6
the PonoMusic Team
LeVar Burton & Reading Rainbow
Altered:
Andromium Inc.
OSSIC
Revols
WobbleWorks LLC.
Cloud Imperium Games Corporation
Mobvoi
Next Thing Co.
Ravean Team
Jide Tech
Rocketbook
Nura
PINE64 Inc.
Shai Goitein
UDOO
LifeBEAM
BLOCKS Wearables
Primo Toys
Creator Name

GDP vs Funding



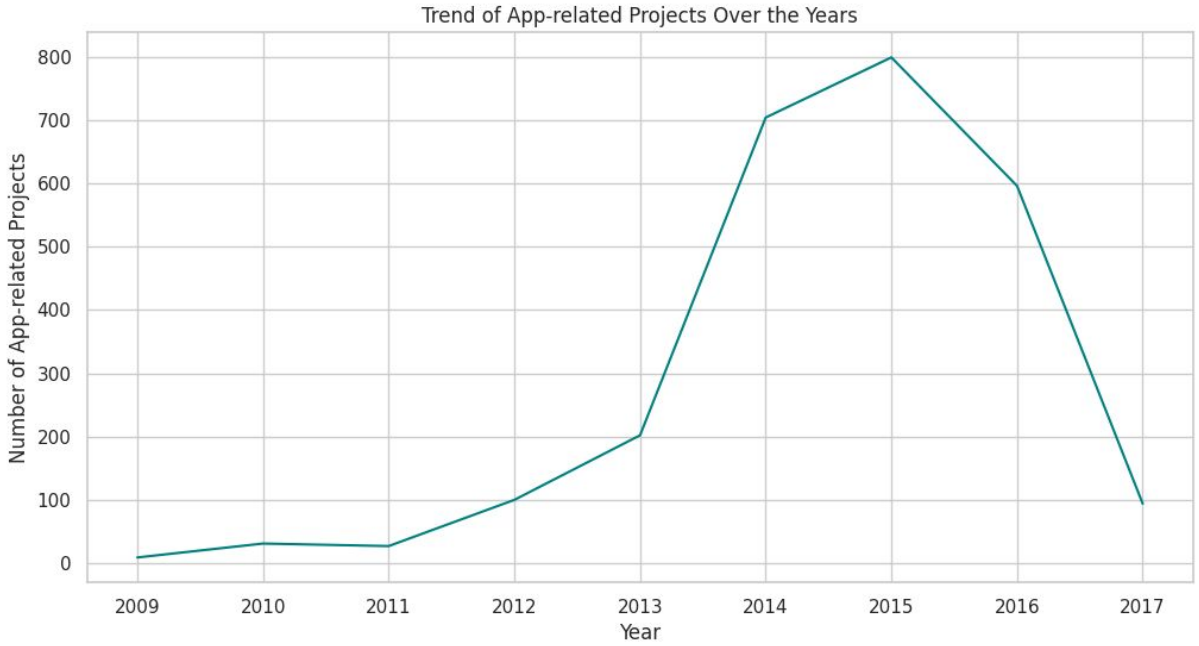
For **2015**, a bar plot of total Kickstarter funding is shown alongside a line plot of GDP (in trillions) for all countries, highlighting the relationship between each country's economic performance and crowdfunding activity.



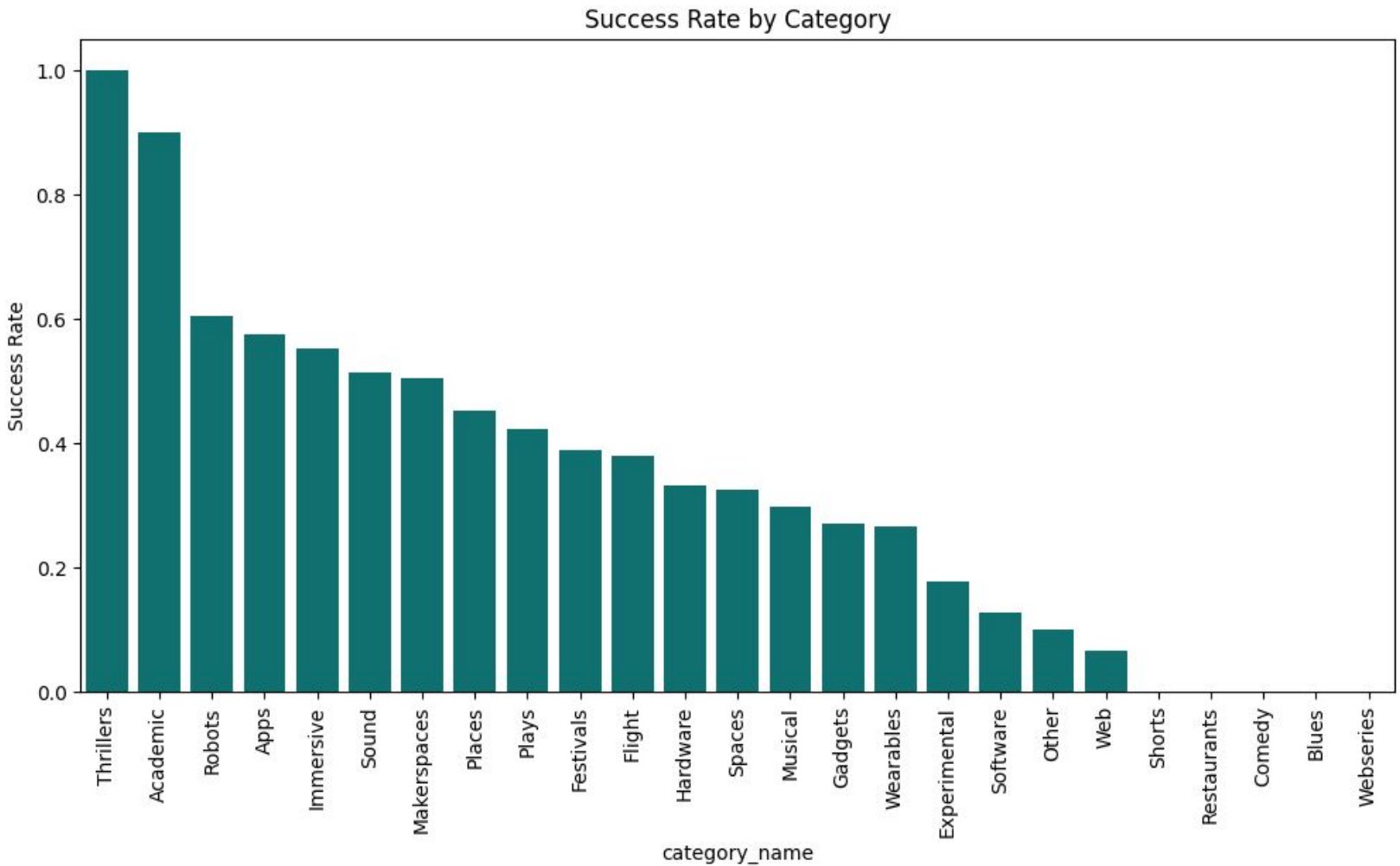
As the GDP increases, we observe a corresponding rise in the 'USD_Pledged' for Kickstarter campaigns in the USA up until 2015. However, there is a noticeable decline in the pledged amounts following **2015**, despite the continued growth in GDP.

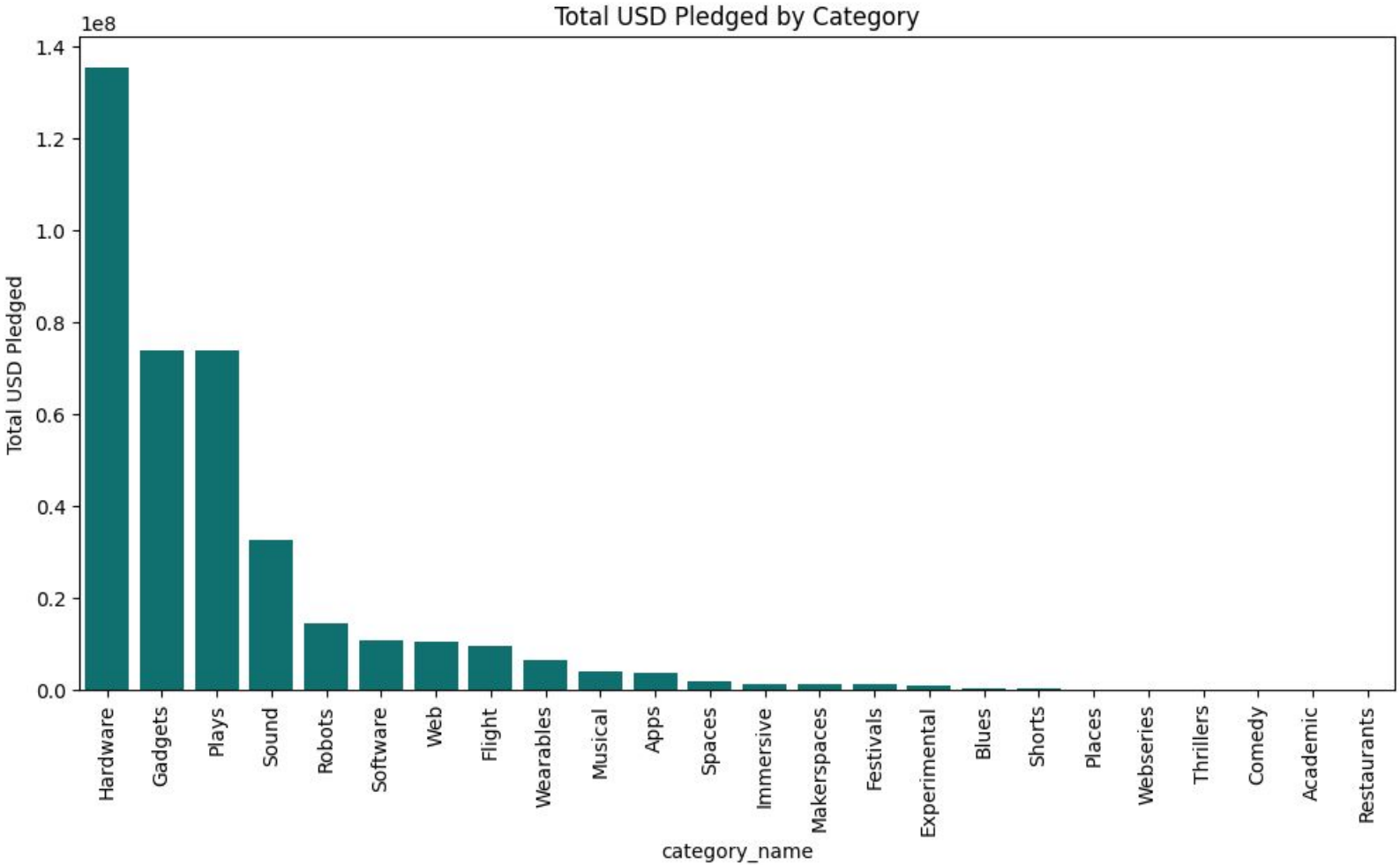
On further investigation it was found that factors like corporate changes and debt flows affected the investment stats.

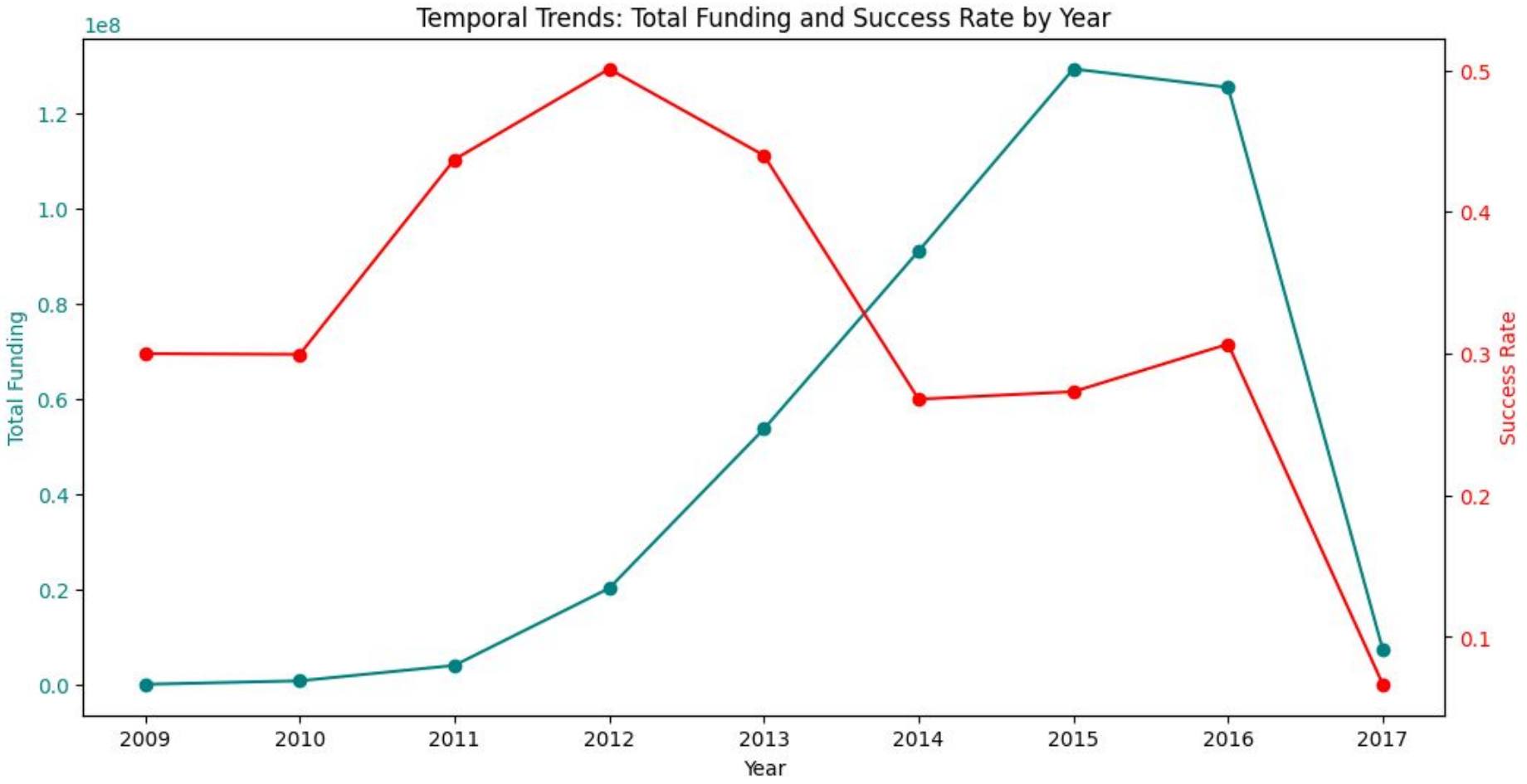
Ref: https://www.federalreserve.gov/econres/notes/feds-notes/what-happened-to-foreign-direct-investment-in-the-united-states-20200213.html

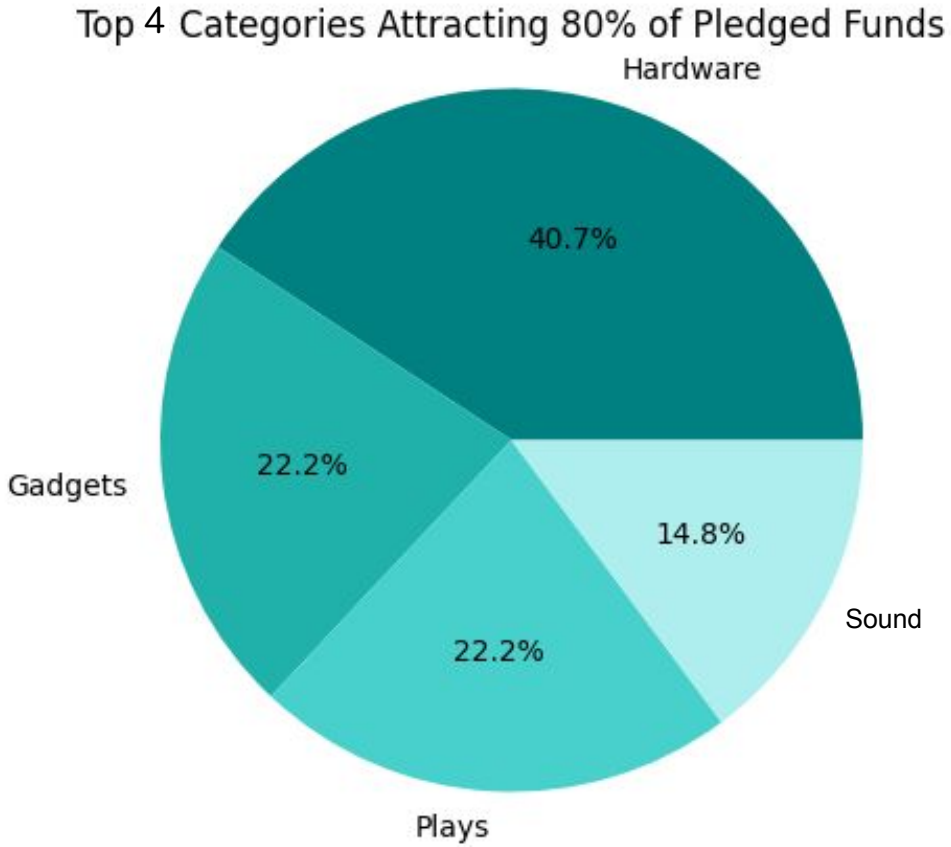


This plot reveals the trend of app-related projects over the years, with a notable rise, beginning in 2013. This aligns with the growing popularity of smartphones and mobile technology.


Success Rate by Category
Success Rate
1.0
0.8
0.6
0.4
0.2
0.0
Thrillers
Academic
Robots
Apps
Immersive
Sound
Makerspaces
Places
Plays
Festivals
Flight
Hardware
Spaces
Musical
Gadgets
Wearables
Experimental
Software
Other
Web
Shorts
Restaurants
Comedy
Blues
Webseries
category_name

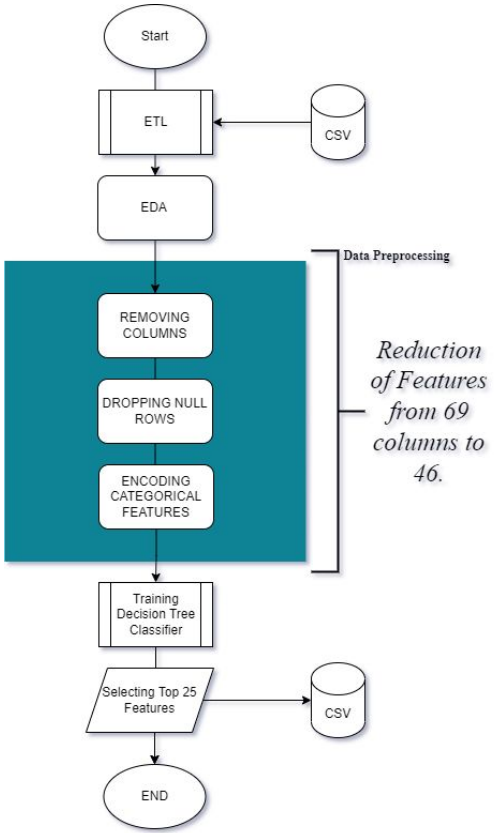

Total USD Pledged by Category
1e8
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Total USD Pledged
Hardware
Gadgets
Plays
Sound
Robots
Software
Web
Flight
Wearables
Musical
Apps
Spaces
Immersive
Makerspaces
Festivals
Experimental
Blues
Shorts
Places
Webseries
Thrillers
Comedy
Academic
Restaurants
category_name


Temporal Trends: Total Funding and Success Rate by Year
1e8
Total Funding
Success Rate
Year
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0.1
0.2
0.3
0.4
0.5
2009
2010
2011
2012
2013
2014
2015
2016
2017


Top 4 Categories Attracting 80% of Pledged Funds
Hardware
40.7%
Gadgets
22.2%
Sound
14.8%
Plays
22.2%

Methodology : Data pre-processing

- Categorical values encoded using label and one-hot encoding
- Reduced dataset from 68 features to 46 after data cleaning
- Identified top 25 features using a decision tree, accounting for 98.84% of the data's importance



Selected Features

Feature	Importance
backers_count	0.2397
launch_to_state_change_days	0.2004
goal	0.1657
pledged	0.1245
launch_to_deadline_days	0.1051
usd_pledged	0.0476
launched_at_yr	0.0394
disable_communication_b	0.0317
blurb_len	0.0026
deadline_month	0.0026
blurb_len_clean	0.0025
launched_at_month	0.0022
state_changed_at_day	0.0022
launched_at_day	0.0022
static_usd_rate	0.0021
created_at_day	0.0020
created_at_hr	0.0020
created_at_month	0.0020
create_to_launch_days	0.0019
name_len	0.0018
state_changed_at_hr	0.0018
founder_name	0.0018
name_len_clean	0.0016
state_changed_at_yr	0.0015
Deadline_yr	0.0014
Total Importance	**98.84**

Methodology：Model Selection

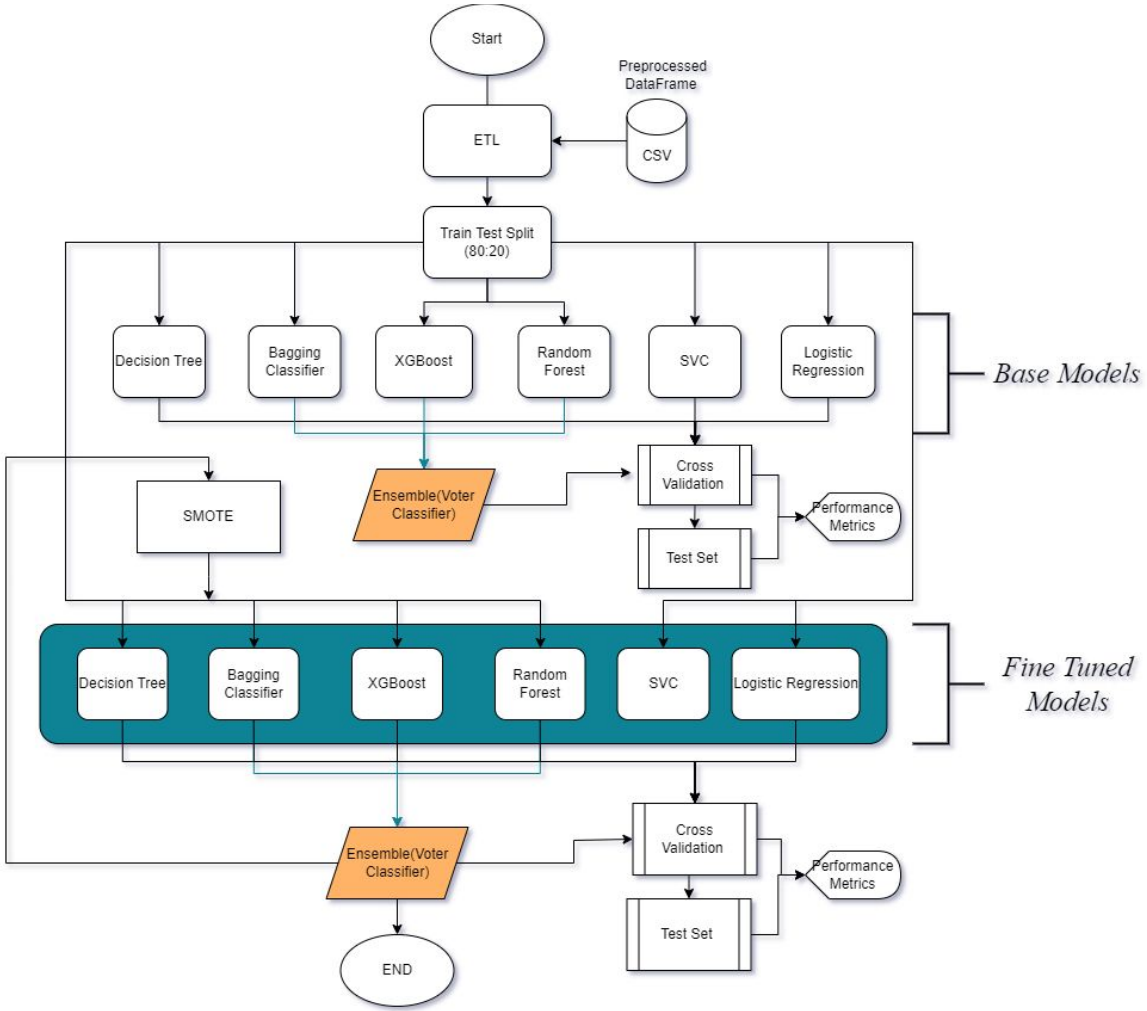
- Models used:

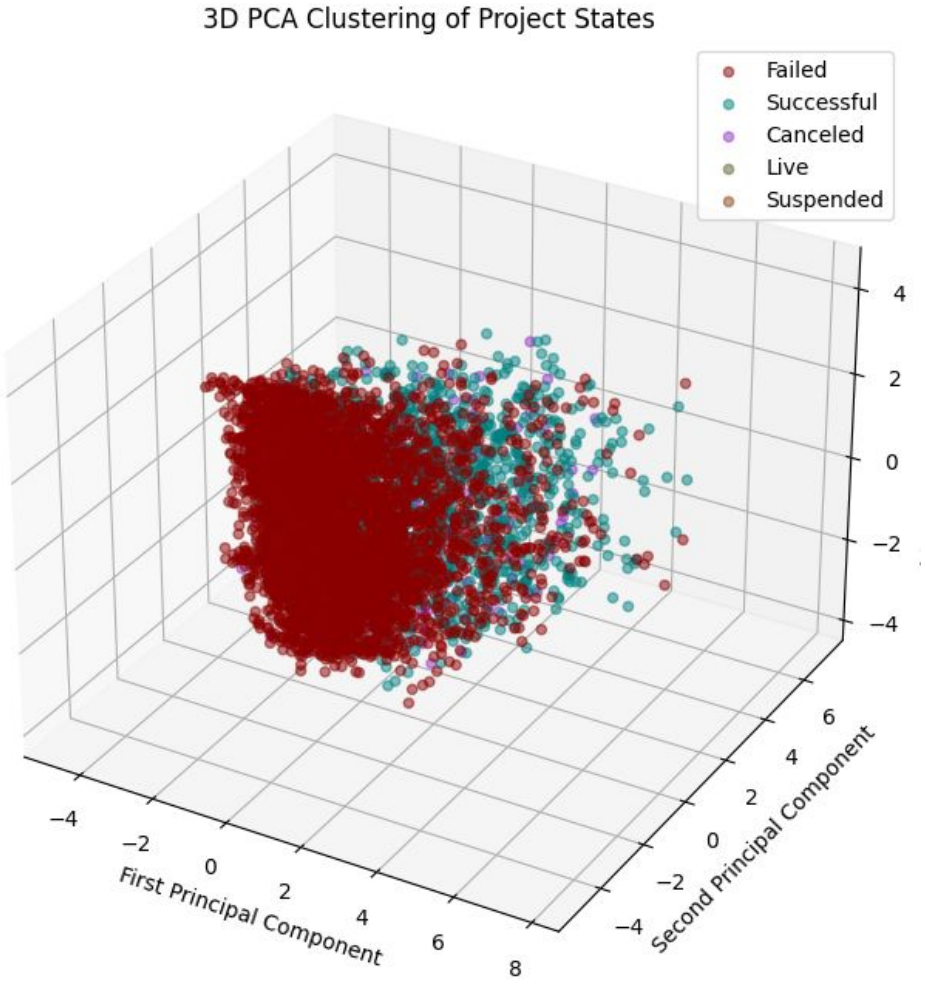
1. Logistic Regression
2. Support Vector Classifier
3. Decision Tree
4. AdaBoost
5. Random Forest
6. Bagging,
7. XGBoost

- Rationale: Chosen models perform well with high-dimensional data

- Data splitting: Stratified sampling used during train-test split to prevent skewed data

- Evaluation metrics: Precision, Recall, F1-Score, and Accuracy


Start
Preprocessed DataFrame
ETL
CSV
Train Test Split (80:20)
Decision Tree
Bagging Classifier
XGBoost
Random Forest
SVC
Logistic Regression
Base Models
SMOTE
Ensemble(Voter Classifier)
Cross Validation
Test Set
Performance Metrics
Decision Tree
Bagging Classifier
XGBoost
Random Forest
SVC
Logistic Regression
Fine Tuned Models
Ensemble(Voter Classifier)
Cross Validation
Test Set
Performance Metrics
END



- Applied PCA on top 25 features to reduce dimensionality
- Extracted 3 principal components
- Created a 3D scatter plot to reveal relationships and patterns in the transformed dataset

Results : Base Classifiers vs Fine-Tuned

Classifiers	Test Accuracy (%)	F1 Score (%)	Scaled Features (Accuracy)	Scaled Features (F1 Score)
Random Forest	96.02 %	95.88 %	----	----
XGBoost	98.61 %	98.41 %	----	----
Decision Tree	95.68 %	93.49 %	----	----
Logistic Regression	78.89 %	34.63 %	84.07%	89.41%
Bagging	97.26%	95.58%	----	----
ADABoost	71.98%	53.46%	----	----
SVM	64.54%	24.95%	78.40%	83.92%

Classifiers	Test Accuracy (%)	F1 Score (%)	Scaled Features (Accuracy)	Scaled Features (F1 Score)
Random Forest	97.81 %	97.92 %	----	----
XGBoost	**98.73 %**	**98.74 %**	----	----
Decision Tree	95.90 %	96.25 %	----	----
Logistic Regression	80.05 %	35.60 %	83.78%	89.35%
Bagging	97.60%	97.84%	----	----
ADABoost	96.02%	96.43%	----	----
SVM	64.44%	24.87%	78.49%	24.87%

Voting Classifier and Performance Metrics

- Combined individual fine tuned classifiers using Voting Classifier

- Enhanced model performance by taking a vote of each classifier and selecting the prediction having highest vote.

- Evaluated the voterclassifier model using performance metrics:

1. Precision
2. Recall
3. F1-Score
4. Accuracy

Voting Classifier and Performance Metrics

Base Model Voter Classifier
(XGB, Bagging, Random Forest)

	precision	recall	f1-score	support
0	1.00	0.87	0.93	491
1	0.98	0.99	0.98	2283
2	0.99	1.00	1.00	102
3	0.98	1.00	0.99	1204
4	1.00	0.98	0.99	46
accuracy			0.98	4126
macro avg	0.99	0.97	0.98	4126
weighted avg	0.98	0.98	0.98	4126

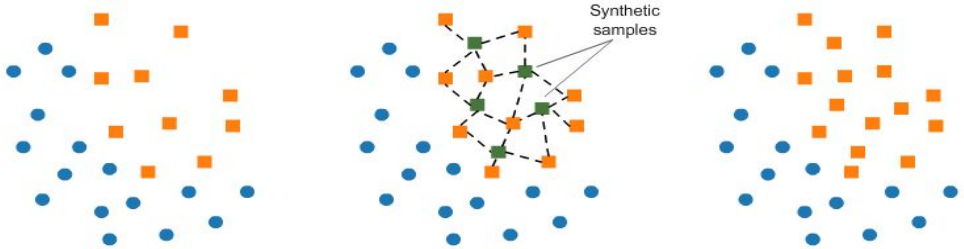
Test Accuracy: 97.93%
F1 Score: 97.89%

Fine-tuned Model Voter Classifier
(XGB, Bagging, Random Forest)

	precision	recall	f1-score	support
0	1.00	0.90	0.95	491
1	0.98	0.99	0.99	2283
2	0.99	1.00	1.00	102
3	0.98	1.00	0.99	1204
4	1.00	1.00	1.00	46
accuracy			0.98	4126
macro avg	0.99	0.98	0.98	4126
weighted avg	0.98	0.98	0.98	4126

Test Accuracy: 98.35%
F1 Score: 98.32%

SMOTE



Classifiers	Test Accuracy (%)	F1 Score (%)
Random Forest	97.81 %	97.92 %
XGBoost	98.73 %	98.74 %
Decision Tree	95.90 %	96.25 %
Logistic Regression	70.74 %	30.53 %
Bagging	97.60%	97.84%
ADABoost	96.02%	96.43%
SVM	64.44%	24.87%

Classifiers with SMOTE	Test Accuracy (%)	F1 Score (%)
Random Forest	97.96 %	98.03 %
XGBoost	98.44%	98.36 %
Decision Tree	95.68 %	93.49 %
Logistic Regression	78.89 %	34.63 %
Bagging	97.60%	97.88%
ADABoost	95.15%	95.66%
SVM	62.57%	31.17%

Thank You