Kickstarter

Amir Ebrahimi - Dat Dang

What is Kickstarter?

Kickstarter is an American public benefit corporation that maintains a global crowdfunding platform focused on creativity. The company's main mission is to help bring creative projects to life!

About the Project

This project is designed and implemented to predict whether or not a project on Kickstarter will be successful based on its features.

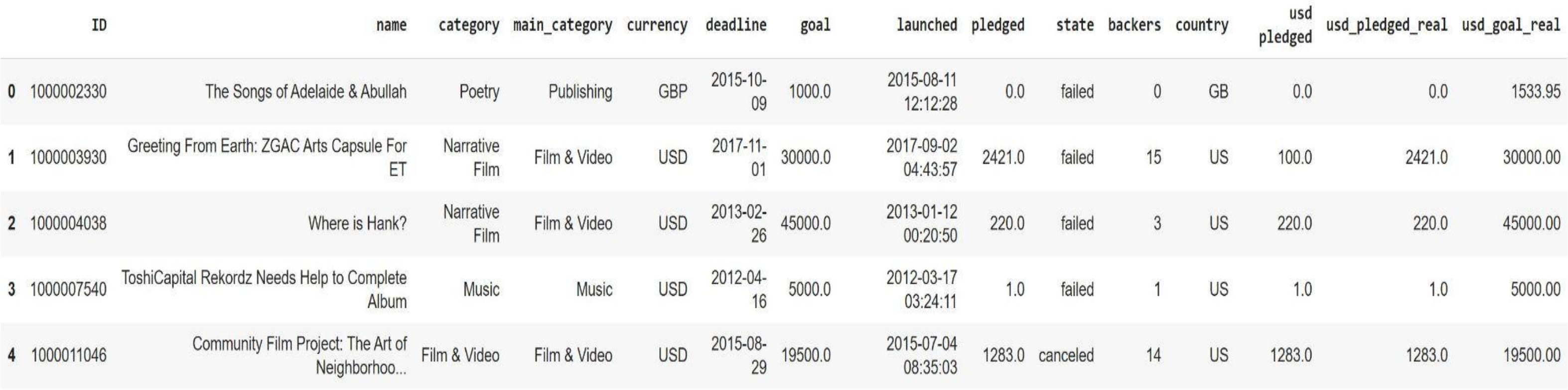
Throughout this presentation of this project, we will discuss how the features are analyzed using different machine learning techniques and algorithms.

About Dataset

The dataset used in this project was accumulated in the year 2018 which includes 702,413 projects with different states that indicate whether a project was successful, failed, canceled, or suspended.

Features included in the dataset are used for the purpose of training and prediction in our machine learning algorithms.

Dataset Features and Details



	ID	name	category	main_category	currency	deadline	goal	launched	pledged	state	backers	country	usd pledged	usd_pledged_real	usd_goal_real
0	1000002330	The Songs of Adelaide & Abullah	Poetry	Publishing	GBP	2015-10-09	1000.0	2015-08-11 12:12:28	0.0	failed	0	GB	0.0	0.0	1533.95
1	1000003930	Greeting From Earth: ZGAC Arts Capsule For ET	Narrative Film	Film & Video	USD	2017-11-01	30000.0	2017-09-02 04:43:57	2421.0	failed	15	US	100.0	2421.0	30000.00
2	1000004038	Where is Hank?	Narrative Film	Film & Video	USD	2013-02-26	45000.0	2013-01-12 00:20:50	220.0	failed	3	US	220.0	220.0	45000.00
3	1000007540	ToshiCapital Rekordz Needs Help to Complete Album	Music	Music	USD	2012-04-16	5000.0	2012-03-17 03:24:11	1.0	failed	1	US	1.0	1.0	5000.00
4	1000011046	Community Film Project: The Art of Neighborhoo...	Film & Video	Film & Video	USD	2015-08-29	19500.0	2015-07-04 08:35:03	1283.0	canceled	14	US	1283.0	1283.0	19500.00

Dataset Labels and Details

The label of the dataset is defined as **State** where it indicates if a project is

- **Successful**
- **Failed**
- **Canceled**
- **Live**
- **Suspended**

Project Goals

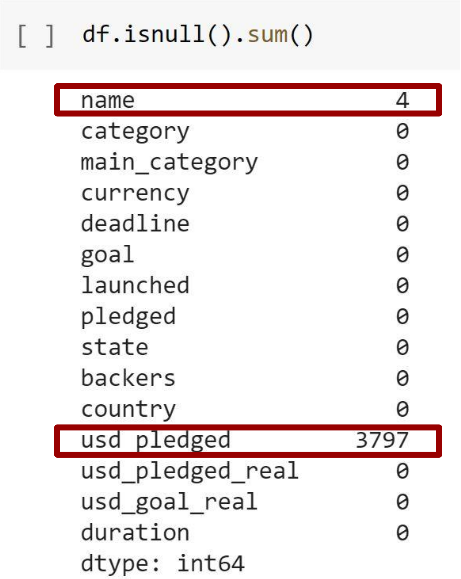
Finding a model that gives the most accurate prediction of whether or not a project has a chance of being successful, using different machine learning techniques we learn in data science and through online research.

Feature Engineering and Data Extraction

To have a better understanding of the dataset and its features, we must extract the most useful part of our raw data. The process of finding the best features requires data analysis and feature engineering which helps us improve the performance of machine learning algorithms.

Feature Engineering and Data Extraction - Cont.

- Finding features that have **null** values

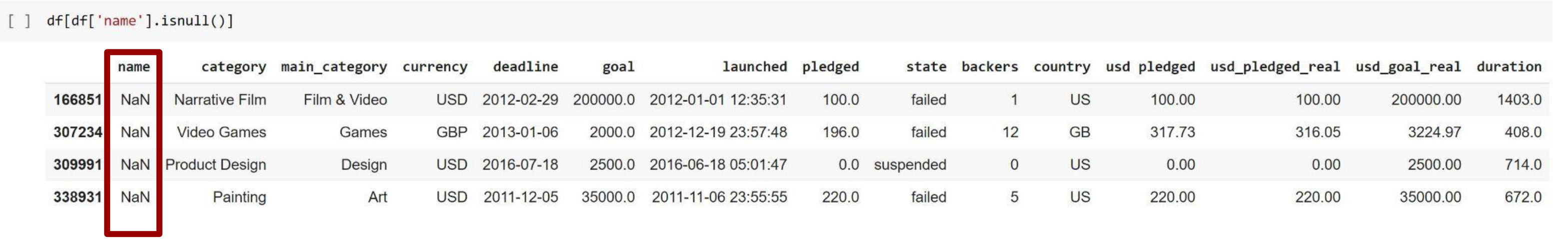


```
[ ]  df.isnull().sum()
```

```
name                 4
category             0
main_category        0
currency             0
deadline             0
goal                 0
launched             0
pledged              0
state                0
backers              0
country              0
usd pledged       3797
usd_pledged_real     0
usd_goal_real        0
duration             0
dtype: int64
```

Feature Engineering and Data Extraction - Cont.

- Exploring projects where they Do Not have a **name**.



```
df[df['name'].isnull()]
```

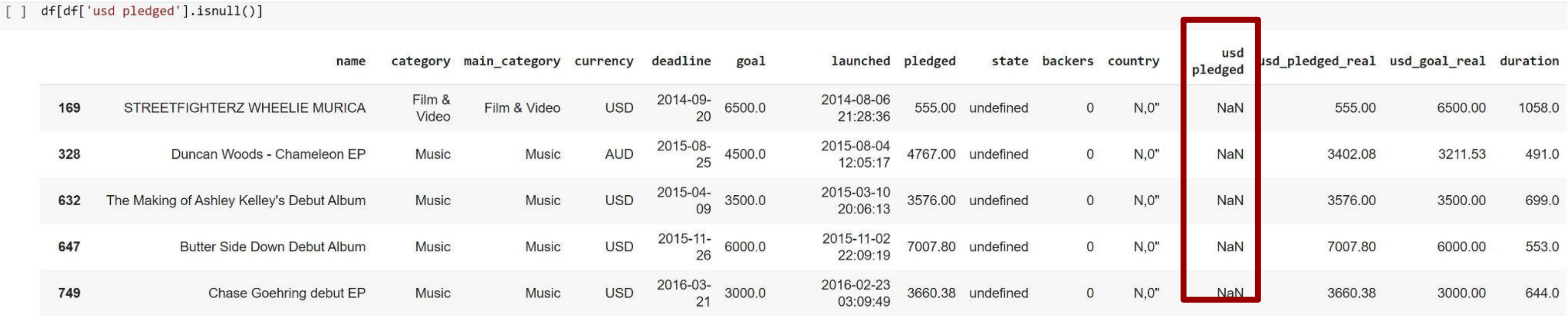
	name	category	main_category	currency	deadline	goal	launched	pledged	state	backers	country	usd pledged	usd_pledged_real	usd_goal_real	duration
166851	NaN	Narrative Film	Film & Video	USD	2012-02-29	200000.0	2012-01-01 12:35:31	100.0	failed	1	US	100.00	100.00	200000.00	1403.0
307234	NaN	Video Games	Games	GBP	2013-01-06	2000.0	2012-12-19 23:57:48	196.0	failed	12	GB	317.73	316.05	3224.97	408.0
309991	NaN	Product Design	Design	USD	2016-07-18	2500.0	2016-06-18 05:01:47	0.0	suspended	0	US	0.00	0.00	2500.00	714.0
338931	NaN	Painting	Art	USD	2011-12-05	35000.0	2011-11-06 23:55:55	220.0	failed	5	US	220.00	220.00	35000.00	672.0

Analyzing:

- Three projects in the U.S. and 1 project in the UK
- Projects with **No Name** most likely be **failed** or **suspended**

Feature Engineering and Data Extraction - Cont.

- Exploring projects where they Do Not **usd pledged**.



```
[ ] df[df['usd pledged'].isnull()]
```

	name	category	main_category	currency	deadline	goal	launched	pledged	state	backers	country	usd pledged	usd_pledged_real	usd_goal_real	duration
169	STREETFIGHTERZ WHEELIE MURICA	Film & Video	Film & Video	USD	2014-09-20	6500.0	2014-08-06 21:28:36	555.00	undefined	0	N,0"	NaN	555.00	6500.00	1058.0
328	Duncan Woods - Chameleon EP	Music	Music	AUD	2015-08-25	4500.0	2015-08-04 12:05:17	4767.00	undefined	0	N,0"	NaN	3402.08	3211.53	491.0
632	The Making of Ashley Kelley's Debut Album	Music	Music	USD	2015-04-09	3500.0	2015-03-10 20:06:13	3576.00	undefined	0	N,0"	NaN	3576.00	3500.00	699.0
647	Butter Side Down Debut Album	Music	Music	USD	2015-11-26	6000.0	2015-11-02 22:09:19	7007.80	undefined	0	N,0"	NaN	7007.80	6000.00	553.0
749	Chase Goehring debut EP	Music	Music	USD	2016-03-21	3000.0	2016-02-23 03:09:49	3660.38	undefined	0	N,0"	NaN	3660.38	3000.00	644.0

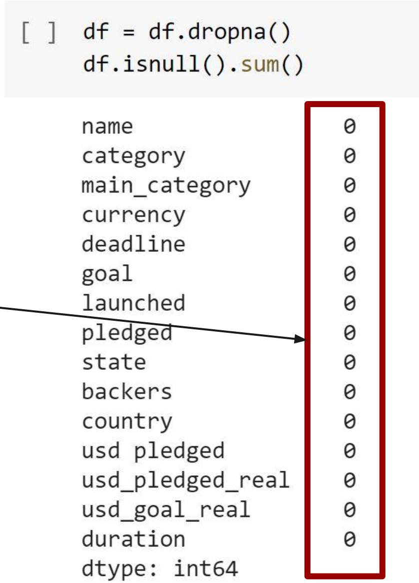
Analyzing:

- ➢ State (Label) cannot be determined (undefined) without knowing how much money is being pledged.

Feature Engineering and Data Extraction - Cont.

- Drop projects from the dataset where they are **NULL**

No **NULL** records in the dataframe

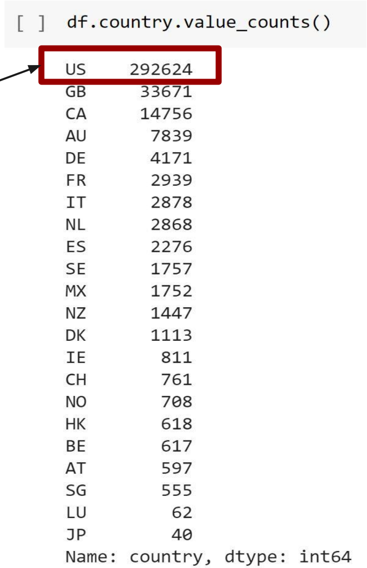


```
[ ]  df = df.dropna()
     df.isnull().sum()
```

```
name                0
category            0
main_category       0
currency            0
deadline            0
goal                0
launched            0
pledged             0
state               0
backers             0
country             0
usd pledged         0
usd_pledged_real    0
usd_goal_real       0
duration            0
dtype: int64
```

Feature Engineering and Data Extraction - Cont.

- Counting the number of projects for each **country**



```
[ ]  df.country.value_counts()
```

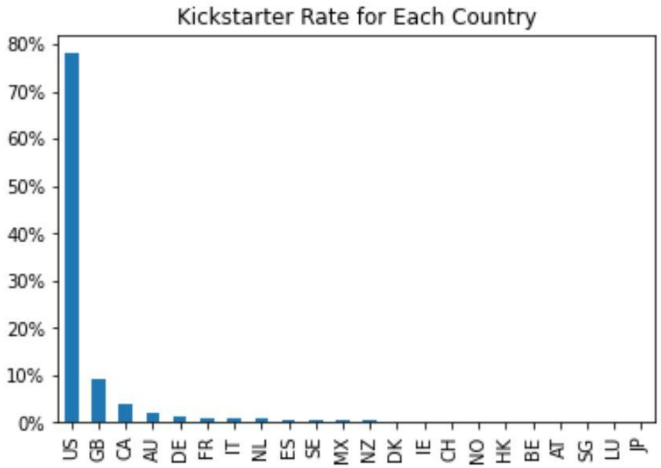
```
US    292624
GB     33671
CA     14756
AU      7839
DE      4171
FR      2939
IT      2878
NL      2868
ES      2276
SE      1757
MX      1752
NZ      1447
DK      1113
IE       811
CH       761
NO       708
HK       618
BE       617
AT       597
SG       555
LU        62
JP        40
Name: country, dtype: int64
```

Note: The U.S. has the most projects in our dataset

Feature Engineering and Data Extraction - Cont.

Notes:

- Since the U.S. Kickstarters constitutes about 78.06% of all Kickstarter projects, we will ONLY implement our machine learning algorithms for the projects based in the **U.S.**
- *Dropping* unnecessary columns

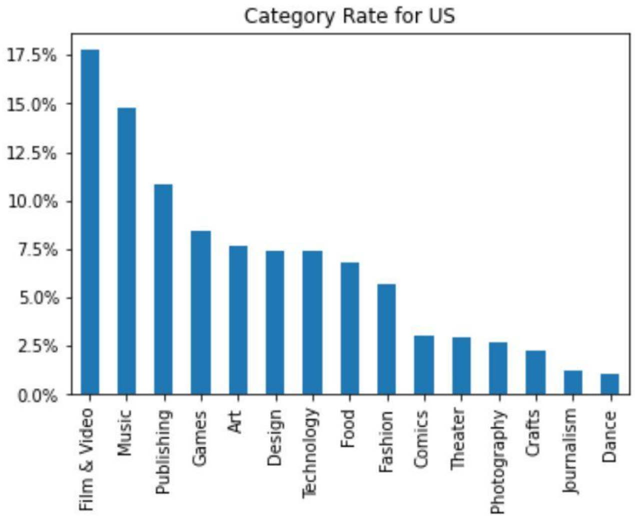


Feature Engineering and Data Extraction - Cont.

- Dropping unnecessary columns (features), such as ID, Name, Country, Currency, category, usd pledged, usd_pledged_real, and usd_goal_real
  - Using **ID** and Project **Name** cannot contribute anything to our machine learning.
  - **Country** and **currency** are useless since we only consider the U.S based Kickstarter projects.
  - **Category** has many different types and does not contribute anything as a feature in the algorithms we used.
  - **usd pledged**, **usd_pledged_real**, and **usd_goal_real** could be useful in the case of the existence of other countries in the dataset.

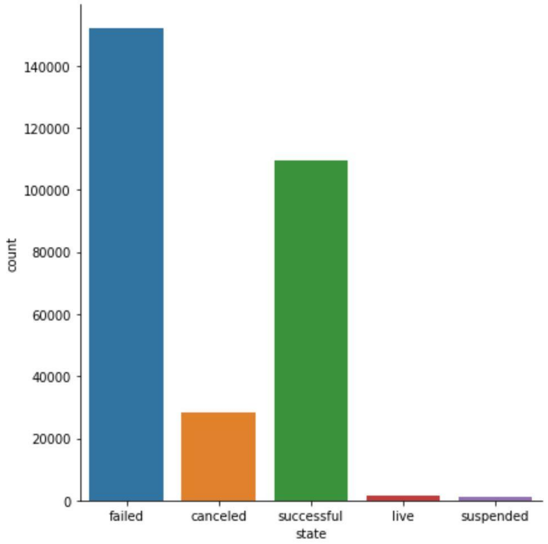
Feature Engineering and Data Extraction - Cont.

- ❖ **Main_Category** is a feature that categorizes Kickstarter projects into fifteen categories.
- ❖ **Film & Video** and **Music** have the majority of the number of Kickstarter projects in the **U.S.**



Feature Engineering and Data Extraction - Cont.

- The **Majority** of Kickstarter projects **fail**.
- Our focus is to predict the status of the projects as **Failed**, **Successful** or **Canceled**.
- Removing **live** and **suspended** from our dataset.

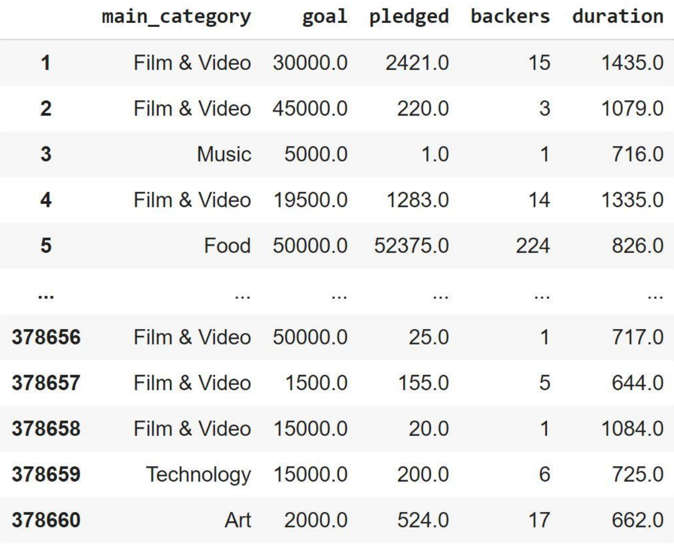


Feature Engineering and Data Extraction - Cont.

- Data Manipulation
  - In order to have a more meaningful feature, the project **deadline** is subtracted by the project **launched** to generate a new and more effective feature named **duration**.
  - After generating **duration**, we dropped both **deadlines** and **launched** from our data frame.

Feature Engineering and Data Extraction - Cont.

- A dataset with more effective features can help increase the performance of our machine learning algorithms!



	main_category	goal	pledged	backers	duration
1	Film & Video	30000.0	2421.0	15	1435.0
2	Film & Video	45000.0	220.0	3	1079.0
3	Music	5000.0	1.0	1	716.0
4	Film & Video	19500.0	1283.0	14	1335.0
5	Food	50000.0	52375.0	224	826.0
...	...	...	...	...	...
378656	Film & Video	50000.0	25.0	1	717.0
378657	Film & Video	1500.0	155.0	5	644.0
378658	Film & Video	15000.0	20.0	1	1084.0
378659	Technology	15000.0	200.0	6	725.0
378660	Art	2000.0	524.0	17	662.0

Machine Learning and Algorithms

- Random Forest
- KNN
- Decision Tree Classifier
- Logistic Regression
- ANN using SKLearn

One-Hot Encoding

- Main_category has fifteen different categories. To classify each category, we implemented One-Hot Encoding, where each category became a column that 0 or 1 determines whether the project is associated with that particular project or not.

Training and Testing Sets

- Using **two** techniques:
  - Splitting data into training set and testing set with testing and training size 20% and 80% respectively. **Random state = 2**
  - k-Fold Cross-Validation

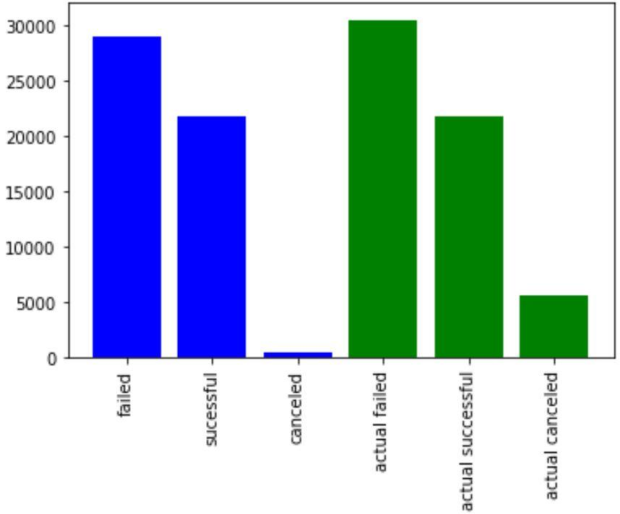
Random Forest - Accuracy and Analysis

- Data Splitting:
  - Accuracy: **88.54%**
- Cross Validation
  - Accuracy: **88.46%**

Analysis

Based on the accuracy and Confusion Matrix, looks like our machine learning cannot predict **canceled** projects and it causes inaccuracy of prediction. However, we have high accuracy in predicting **successful** and **failed** projects.

- Confusion Matrix



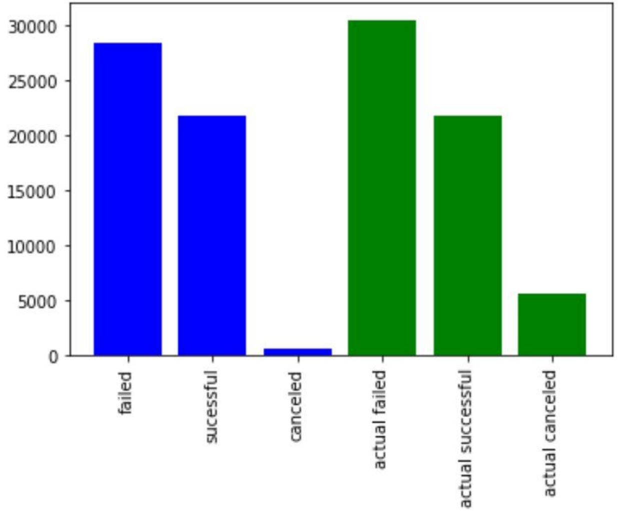
KNN - Accuracy and Analysis

- Data Splitting:
  - Accuracy: **87.67%**
- Cross Validation
  - Accuracy: **87.46%**

Analysis

Lower accuracy compared to Random Forest as our algorithm cannot accurately predict canceled projects. More accurate on predicting successful project.

- Confusion Matrix



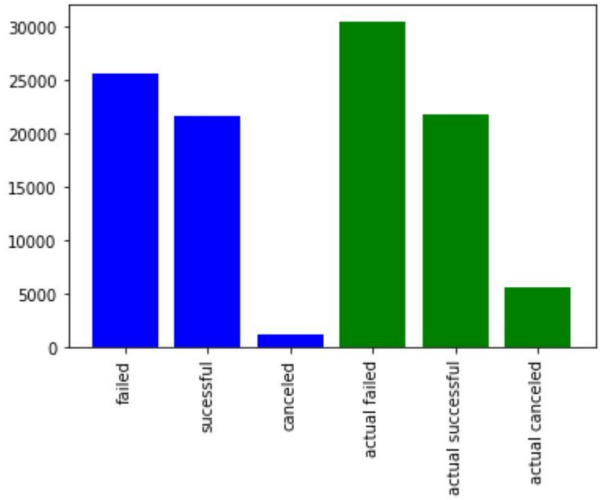
Decision Tree Classifier - Accuracy and Analysis

- Data Splitting:
  - Accuracy: **83.68%**
- Cross Validation
  - Accuracy: **83.57%**

Analysis

Lower accuracy compared to Random Forest and KNN in predicting **failed** projects, however, it predicts more **canceled** projects compared to all other ML algorithms.

- Confusion Matrix



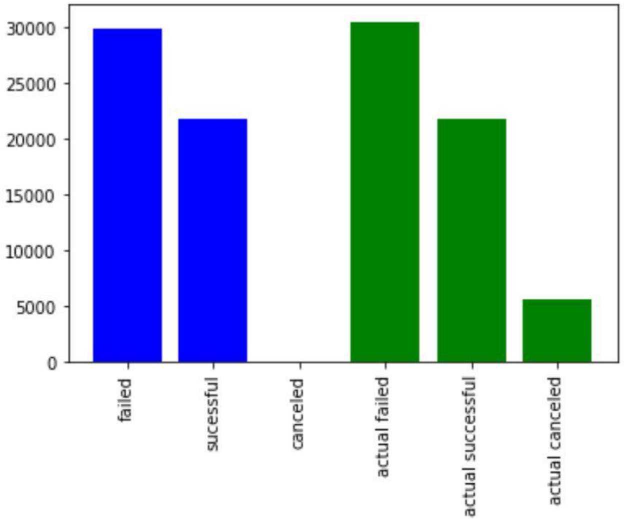
Logistic Regression - Accuracy and Analysis

- Data Splitting:
  - Accuracy: **89.38%**
- Cross Validation
  - Accuracy: **89.40%**

Analysis

Highest accuracy compared to other ML algorithms. About 90% predict **failed** and **successful** projects; whereas, it cannot predict any **canceled** projects.

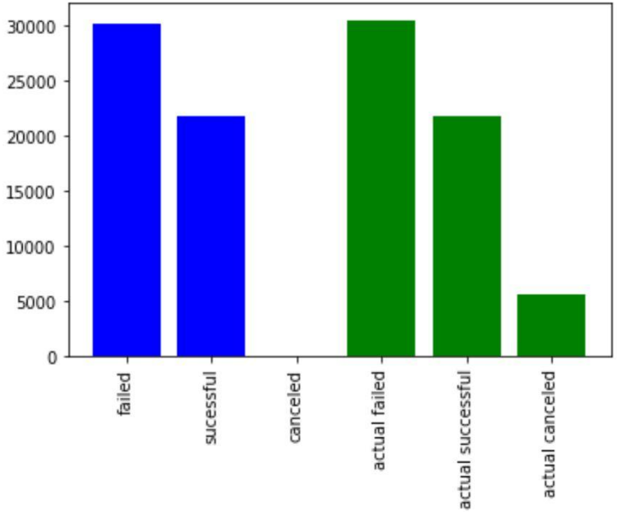
- Confusion Matrix



Artificial Neural Network - Accuracy and Analysis

- Data Splitting:
  - Accuracy: **89.75%**

- Confusion Matrix



Analysis

Similar to Logistic Regression, it is highly accurate compared to other ML algorithms. ANN is slightly more accurate than LR however, it cannot predict **canceled** projects.

Conclusion

- **Highest Accuracy:** ANN has the highest accuracy (**89.75%**) in prediction compared to Random Forest, KNN, Decision Tree, and Logistic Regression.
- **Lowest Accuracy:** The Decision Tree classifier has the lowest accuracy (**83.57%**) compared to other ML algorithms.
- **Common Errors:** We noticed that all algorithms have trouble predicting **Canceled** projects. After analyzing the data and focusing on only **Canceled** projects, we found that projects that are canceled Do Not have logical reasons for cancellation based on their features. For instance, a project can be canceled for a personal reason even if it has many backers and a high amount of pledge. Therefore, our machine learning algorithms cannot accurately predict **canceled** projects!

References

- Kickstarter (Kaggle)