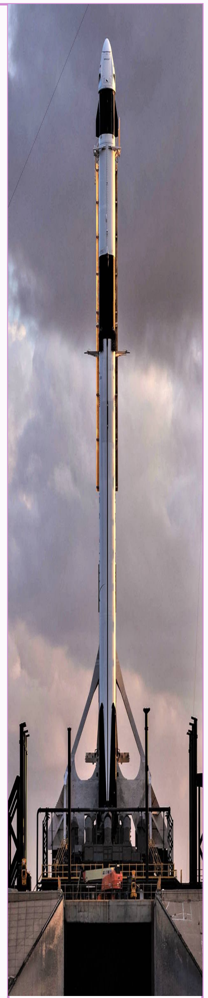
IBM Applied Data Science Capstone

Space X Falcon 9 First stage Landing Prediction

2010- 2020

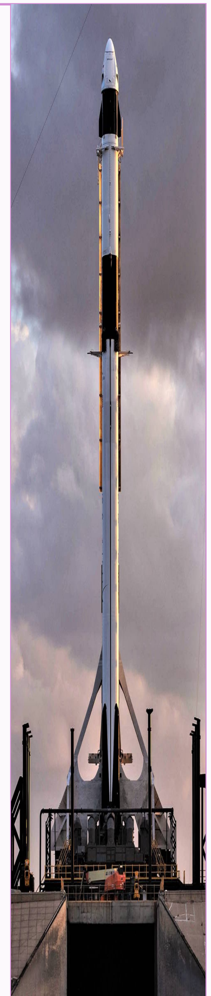
Outline



Executive Summary

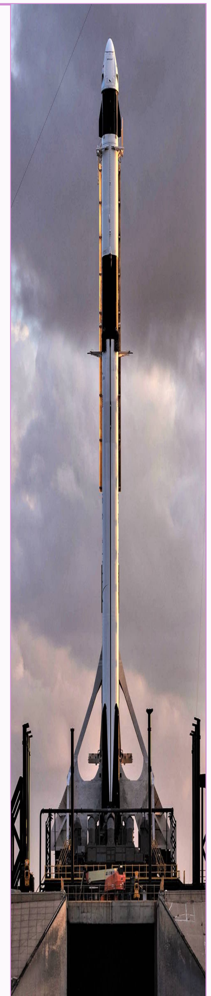
For Space X Falcon 9 (Orbital Rocket capable of Reflight), we want to predict whether the first stage will land properly and therefore be reused.

Best recommandation is to use the Logistic Regression Model for prediction. This model has an accuracy of 82% on the test set.



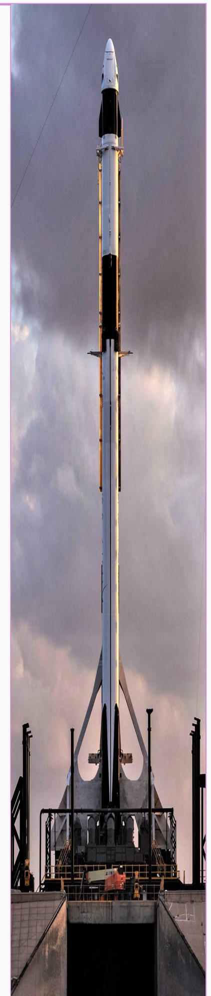
Methodology

- We collect data from a static response object (see Appendix).
- We keep data relating to Falcon 9 only and for the years 2010 to 2020. We finished with **90** records.
- For 5 missing values in the PayLoadMass, we replace the missing values with the average value.
- Based on the dataset with 12 numeric features and 8 categorical features, we conduct some exploratory data analysis to detect variables that can predict the landing results (see Results ..)
- We then select the features for predictions.



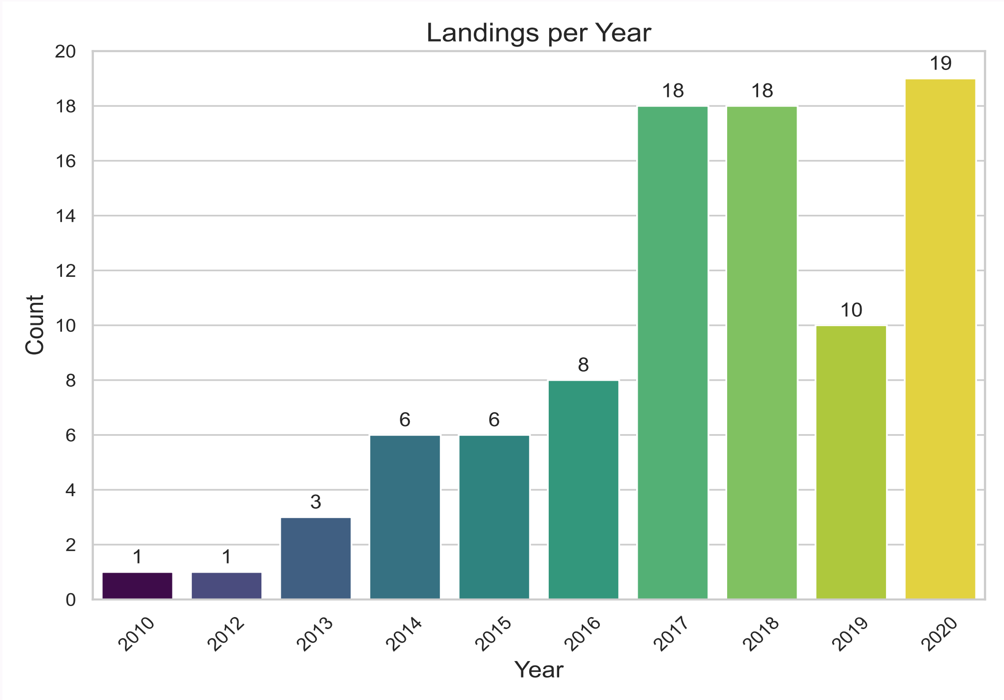
Results Numerical Features

- Number of landings per Year
- Correlation Matrix
- Percentage of Landing Success per legs opening
- Percentage of Landing Success per GridFins
- Percentage of Landing Success per Reused Count
- Percentage of Landing Success per Block
- Percentage of Landing Success per Year
- Percentage of Landing Success per Payload Mass

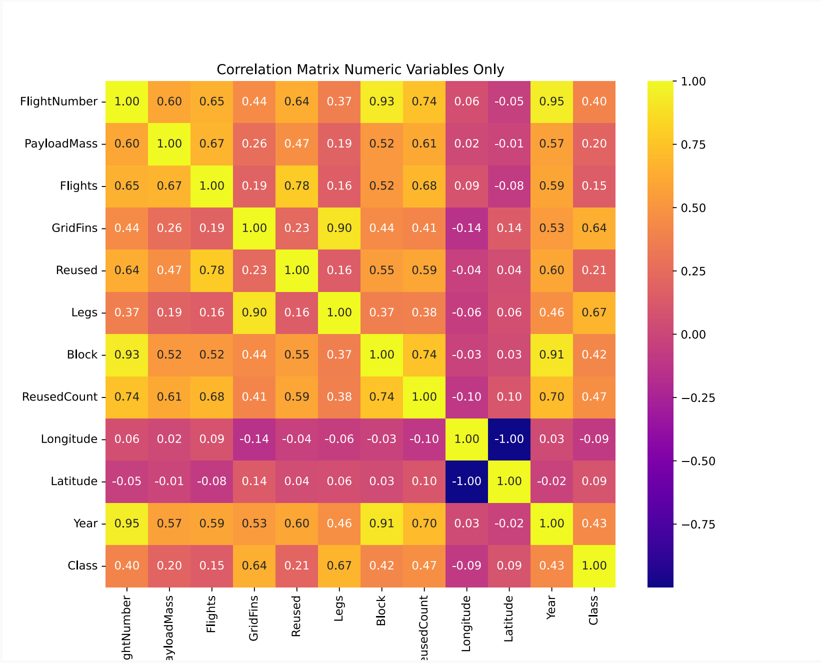


Number of Landings per Year

What are we talking about?

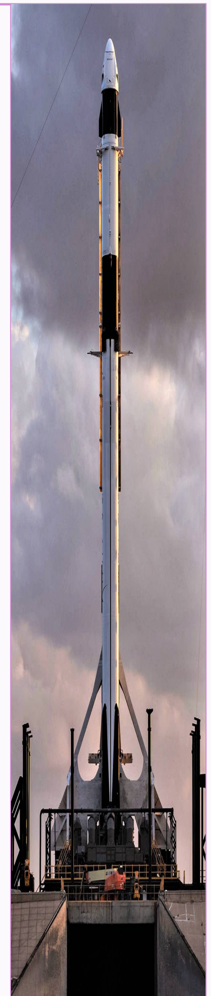


Correlation Matrix (numeric variables)



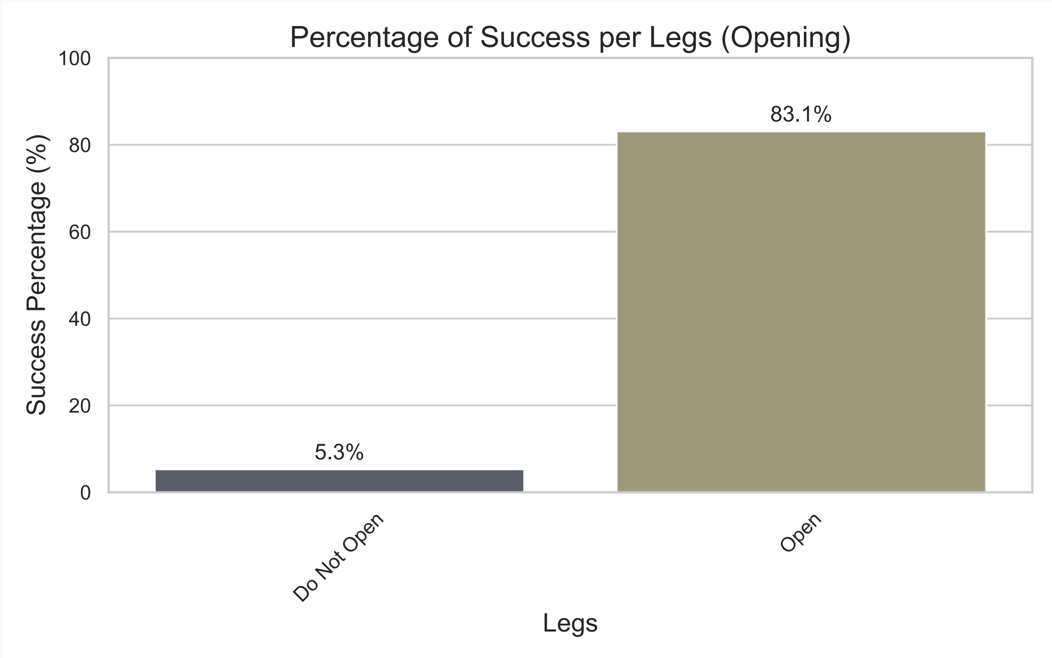
Using a correlation matrix, help viewing how strongly the numeric variables are related to the target value ('Class'):

Legs: 0.67, Gridfins: 0.64,
ReusedCount: 0.47, Year: 0.43
Block: 0.42

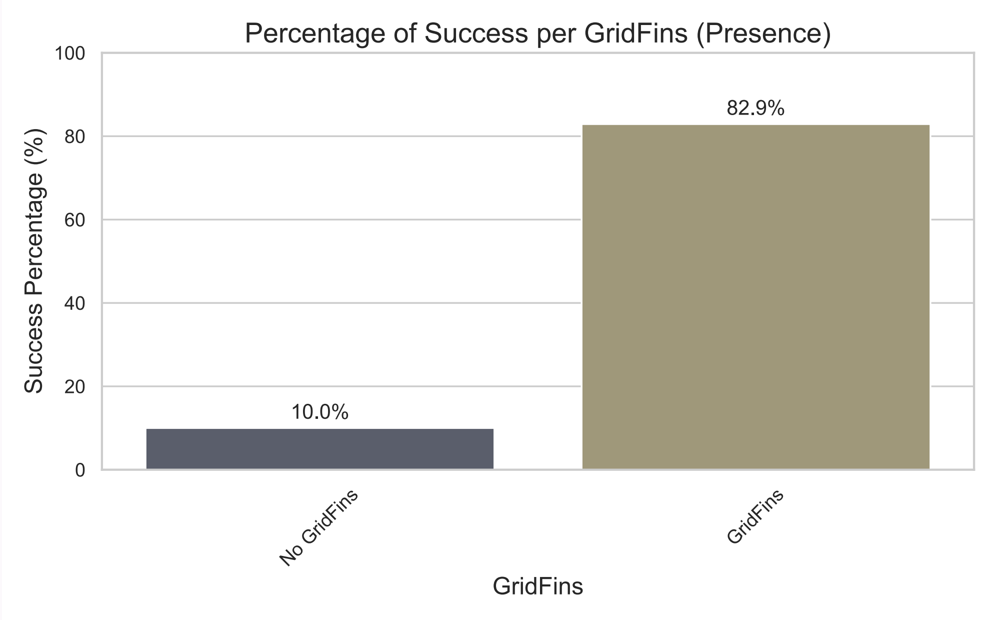


Percentage of Landing Success per legs opening

Quite logical, if the legs do not open then the percentage of success is very low.

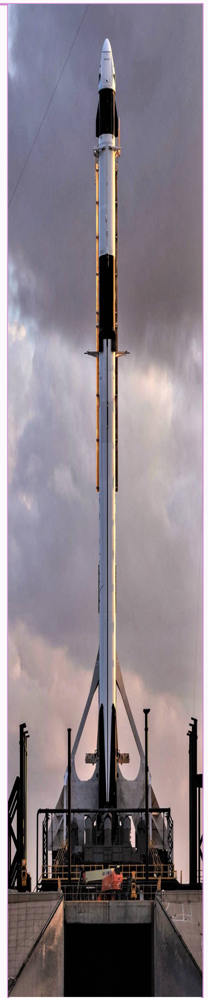


Percentage of Landing success per GridFins



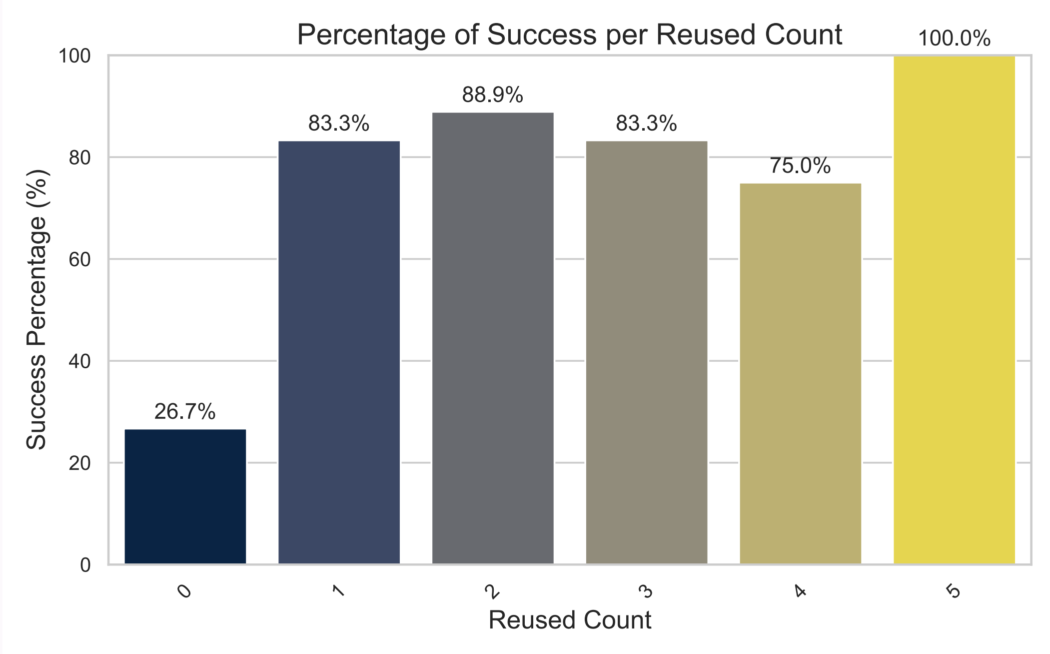
Grid Fins is a type of flight control surface used on rockets.

In case of presence, the percentage of success is much higher.



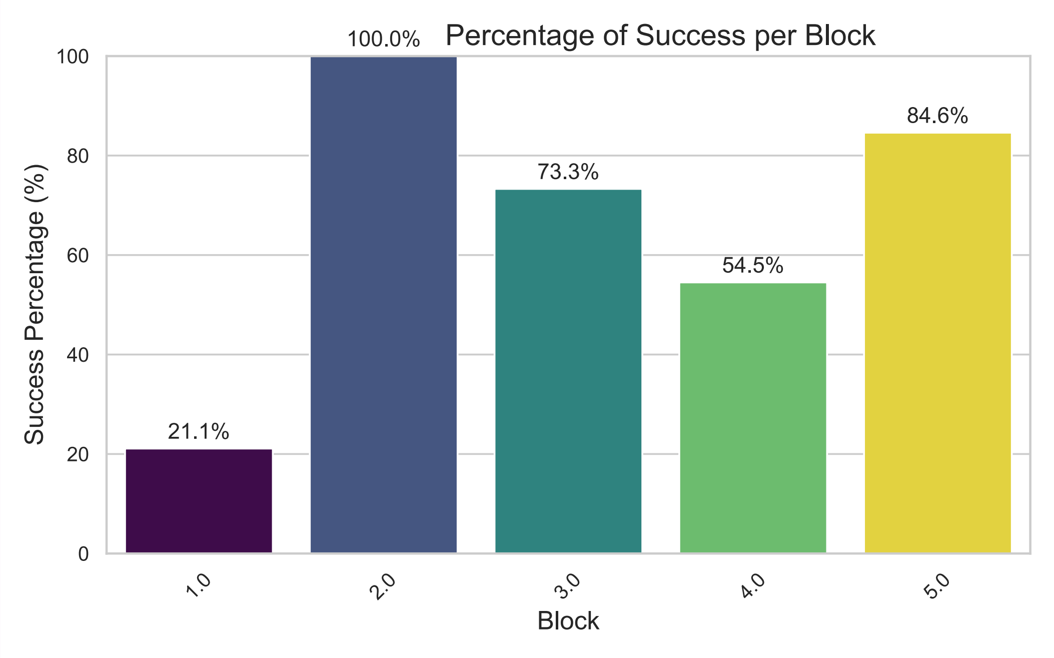
Percentage of Landing success per Reused Count

Excellent performance, in case of reuse the success percentage reaches an average of 87%.

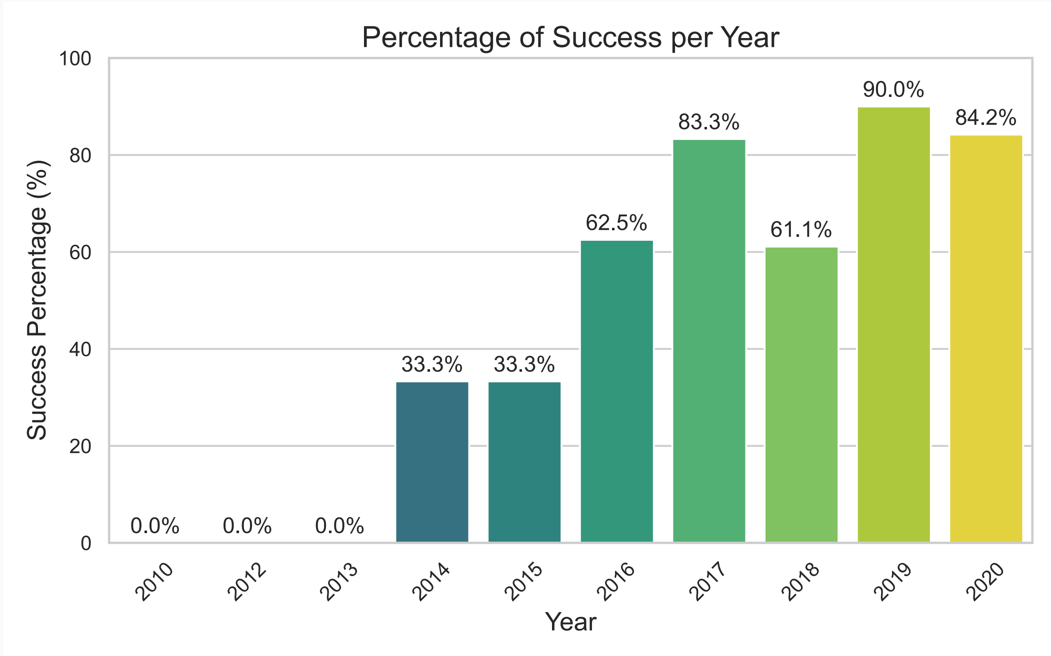


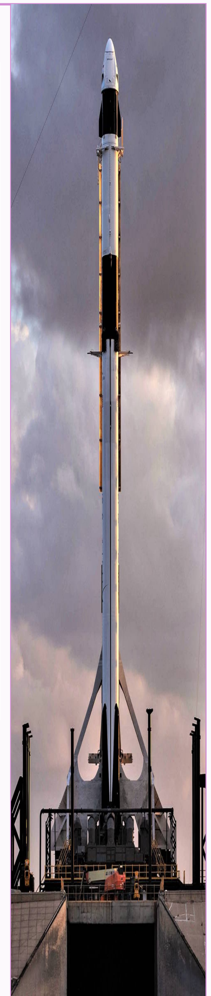
Percentage of Landing success per Block

Blocks are different version of Falcon 9.
Given the block, the rate of success is quite different.



Percentage of Landing success per Year

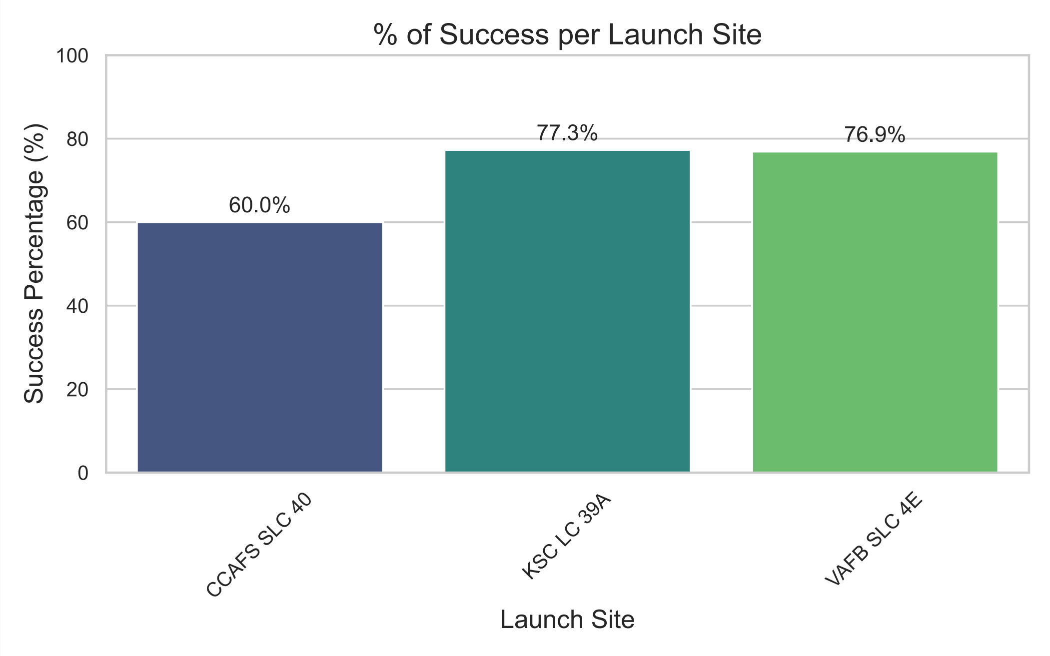




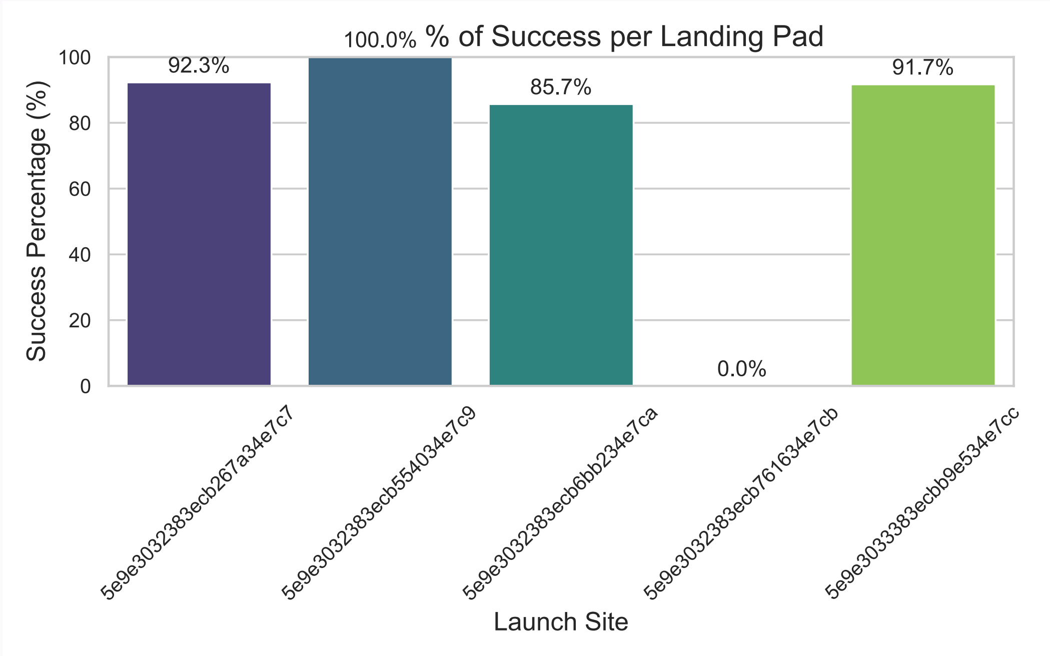
With the help of experience, the success rate is on a upward trend

Percentage of success per Launch Site

At least 60% of success and up to 77%



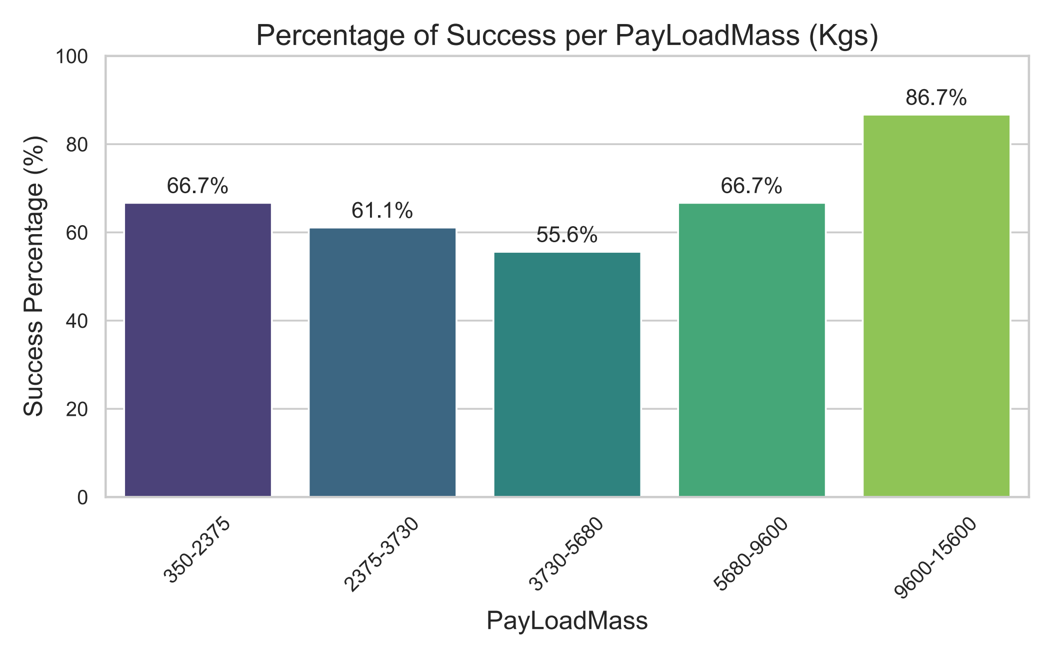
Percentage of success per Landing Pad



Results are consistent.
For the Landing Pad with no success, it concern only 2 landings.

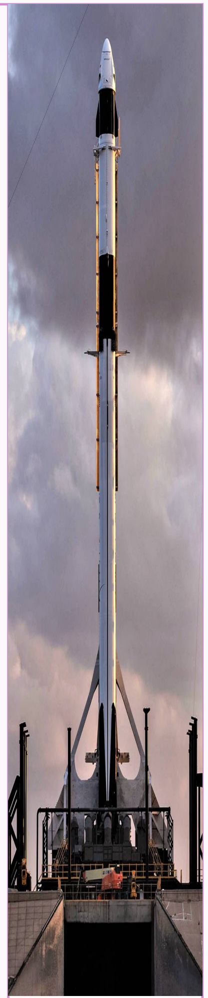
Percentage of success per Payload Mass

Results are somehow different for different carrying capacities.



Results Categorical Features

- Count of unique values per categorical feature
- Percentage of Landing Success per orbit
- Percentage of Landing Success per Launch Site
- Percentage of Landing Success per Landing Pad

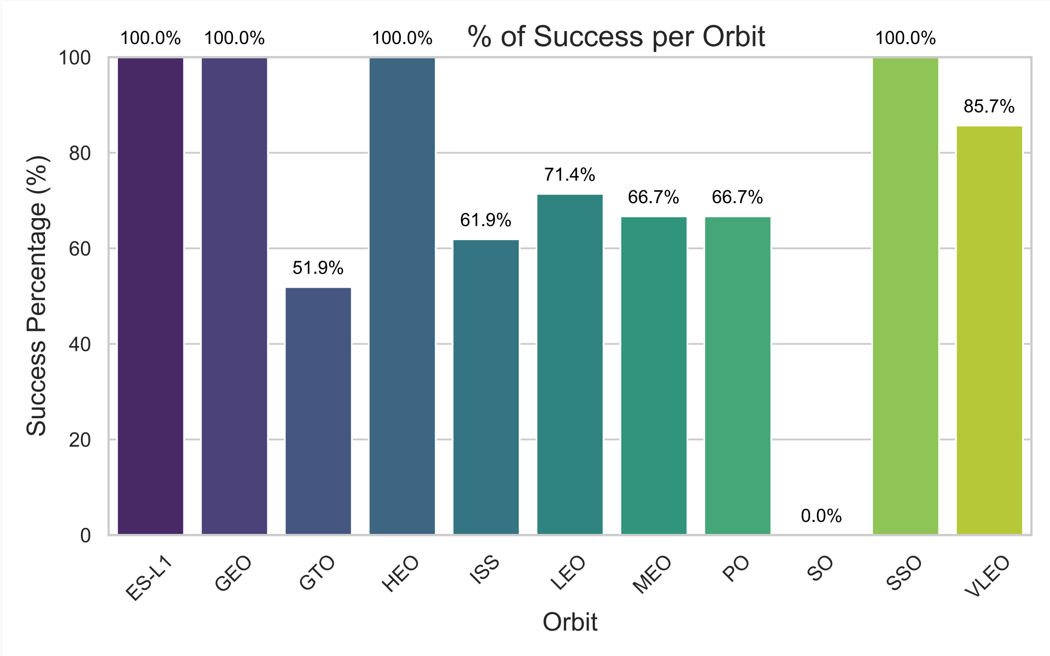


Count of unique value per categorical features

With only 1 occurrence BoosterVersion is not of interest, Serial has too many Unique Count (53) and Outcome is without interest.

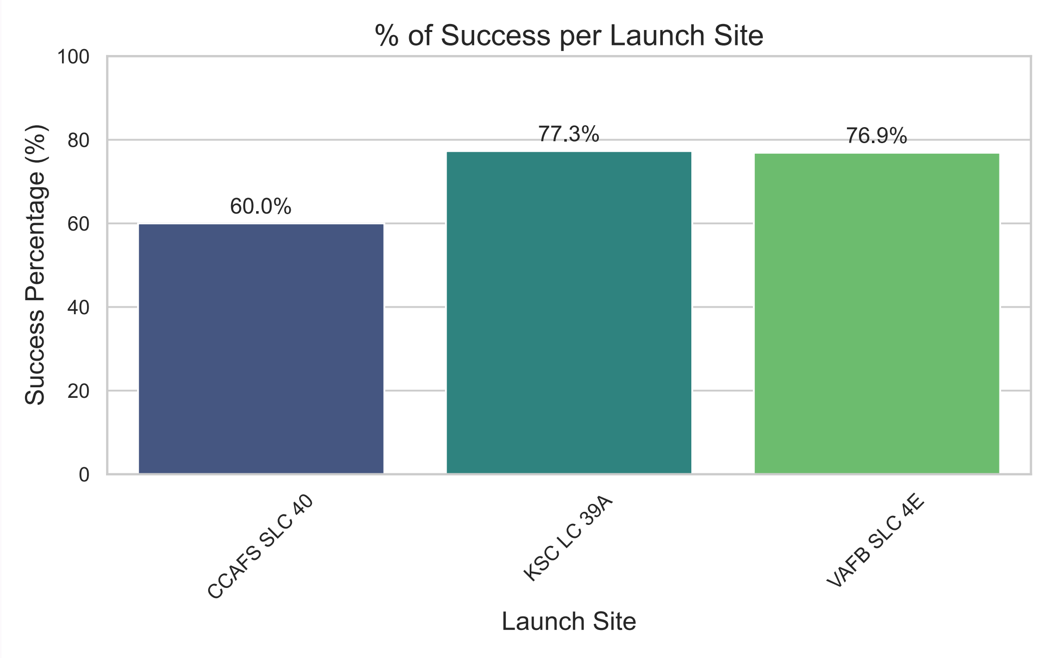
Feature	Unique Count
BoosterVersion	1
Orbit	11
LaunchSite	3
LandingPad	5
Serial	53
Outcome	8

Percentage of Landing Success per Orbit



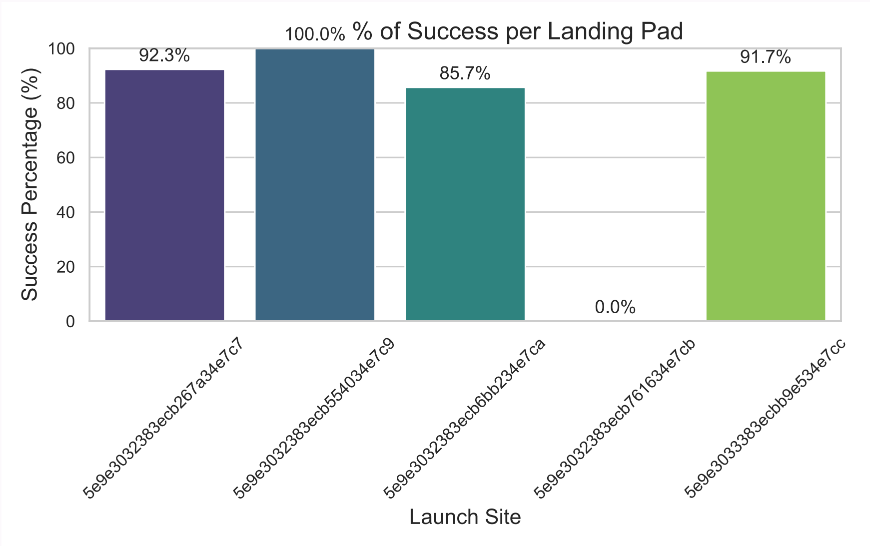
11 orbits and the result is somehow different.

Percentage of Landing Success per Launch Site



A small weakness on the CCAFS site. For the other 2, the success percentage is similar.

Percentage of Landing Success per Landing Pad



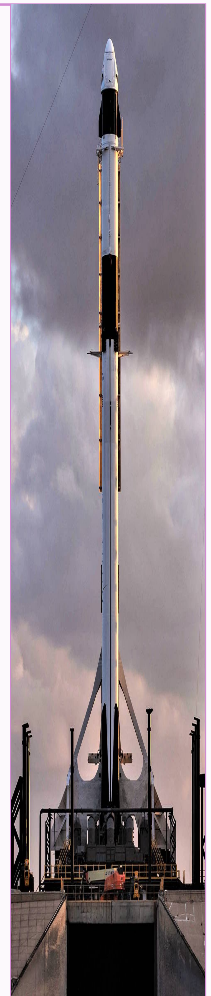
Except for one landing pad (2 landings only), the rate of success is in the same range.

Prediction

Based on the exploratory data analysis, the following data have been retained for the training of the various model:

- Categorical : Orbit , LaunchSite, LandingPad, Payload Mass
- Boolean: Grid Fins, Legs
- Numeric: Block Reused Count Year.

For categorical data, one hot encoding (allows categorical data to be effectively utilized in numeric model) has been applied.

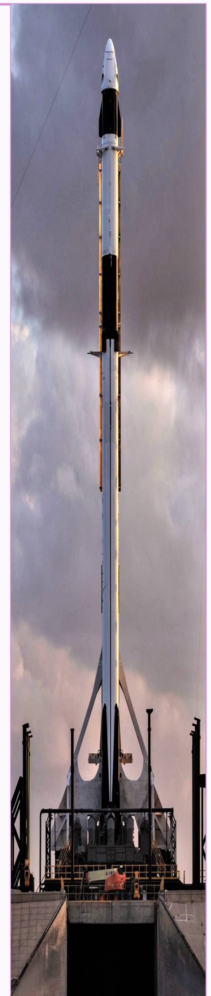


Prediction

For numerical , scaling (normalize the range of features) has been applied.

The dataset has been divided between train dataset and test dataset . (80/20)

Different model have been applied.

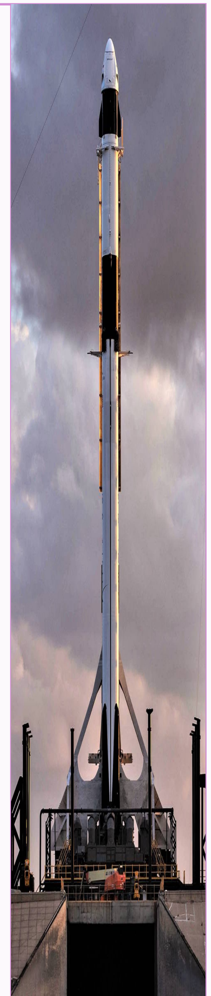


Prediction

The different models can be summarized as follows:

F1 is a balance between Precision and Recall

	LogReg	SVM	Tree	KNN
Precision	0.769231	0.769231	0.818182	0.714286
Recall	1.000000	1.000000	0.900000	1.000000
F1	0.869565	0.869565	0.857143	0.833333
Accuracy	0.823529	0.823529	0.823529	0.764706



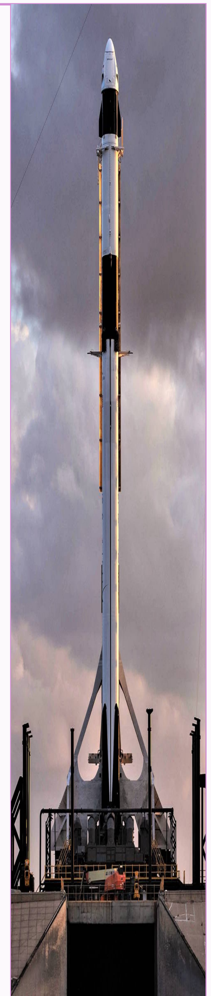
Conclusion

As a reminder, the basic accuracy to beat is 0.71.

All models slightly beat this basic accuracy.

As a matter of fact, we face in this case with datasets of small size (Train: 68 examples, Test: 17 examples).

The most satisfactory model is that of logistic regression (LogReg). This model is often used due to its effectiveness, interpretability, and efficiency.



Appendix

Data Collection

So that we all work on common data, we make use of a static URL https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/dataset_part_2.csv

Python Libraries

Pandas, numpy, request, datetime,