IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Shaojie
12 Dec 2024

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

Exploratory Analysis methods where used in determining the features affecting the success rate of a successful first stage landing of the falcon 9

Explored relationships between

- Flight number and payload mass
- Flight number and launch site
- Payload mass and Launch site
- Orbit type
- Flight number and Orbit type
- Payload Mass and Orbit type

Out of 4 models that we train through **Predictive Classification Analysis**, none yield significant better results

Introduction

Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage.

Therefore **if we can determine if the first stage will land, we can determine the cost of a launch.** This information can be used if an alternate company wants to bid against space X for a rocket launch.

Section
1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected directly from SpaceX website and also scraped from wikipedia
- Perform data wrangling
  - Remove null values, add labels, combine data from different sources, group and average success rate based on launch sites
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Build, tune, evaluate classification models

Data Collection

Data were collected directly via SpaceX api and also scraped from wikipedia

SpaceX APi :

https://api.spacexdata.com/v4/

Wikipedia :
https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

Data Collection – SpaceX API

	static_fire_date_utc	static_fire_date_unix	tbd	net	window	rocket	success	details	cr	crew	ships	capsules	payloads	launchpad	auto_update	failures	flight_number	date_utc	date_unix	date_local	date_precision	upcoming	cores	
0	2006-03-17T00:00:00.000Z	1.142554e+09	False	False	0.0	5e9d0d95eda69955f709d1eb	False	Engine failure at 33 seconds and loss of vehicle		[]	[]	[]	[5eb0e4b5b6c3bb0006eeb1e1]	5e9e4502f5090995de566f86	True	[{'time': 33, 'altitude': None, 'reason': 'merlin engine failure'}]	1	2006-03-24T22:30:00.000Z	1143239400	2006-03-25T10:30:00+12:00	hour	False	[{'core': '5e9e289df35918033d3b2623', 'flight': 1, 'gridfins': False, 'legs': False, 'reused': False, 'landing_attempt': False, 'landing_success': None, 'landing_type': None, 'landpad': None}]	5eb87cd9ffd86e
1	None	NaN	False	False	0.0	5e9d0d95eda69955f709d1eb	False	Successful first stage burn and transition to second stage, maximum altitude 289 km, Premature engine shutdown at T+7 min 30 s, Failed to reach orbit, Failed to recover first stage		[]	[]	[]	[5eb0e4b6b6c3bb0006eeb1e2]	5e9e4502f5090995de566f86	True	[{'time': 301, 'altitude': 289, 'reason': 'harmonic oscillation leading to premature engine shutdown'}]	2	2007-03-21T01:10:00.000Z	1174439400	2007-03-21T13:10:00+12:00	hour	False	[{'core': '5e9e289ef35918416a3b2624', 'flight': 1, 'gridfins': False, 'legs': False, 'reused': False, 'landing_attempt': False, 'landing_success': None, 'landing_type': None, 'landpad': None}]	5eb87cdaffd86e

Access the data collection api source file here:

https://github.com/crayoncannon/IBM-Data-Science-Professional-Certificate/blob/main/Final_Project/1.%20Data%20Collection%20API.ipynb

Data Collection - Scraping

[hide] Flight No.	Date and time (UTC)	Version, Booster[b]	Launch site	Payload[c]	Payload mass	Orbit	Customer	Launch outcome	Booster landing
78	7 January 2020, 02:19:21[492]	F9 B5 B1049.4	CCAFS, SLC-40	Starlink 2 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]	LEO	SpaceX	Success	Success (drone ship)
	Third large batch and second operational flight of Starlink constellation. One of the 60 satellites included a test coating to make the satellite less reflective, and thus less likely to interfere with ground-based astronomical observations.[493]								
79	19 January 2020, 15:30[494]	F9 B5 B1046.4	KSC, LC-39A	Crew Dragon in-flight abort test[495] (Dragon C205.1)	12,050 kg (26,570 lb)	Sub-orbital[496]	NASA (CTS)[497]	Success	No attempt
	An atmospheric test of the Dragon 2 abort system after Max Q. The capsule fired its SuperDraco engines, reached an apogee of 40 km (25 mi), deployed parachutes after reentry, and splashed down in the ocean 31 km (19 mi) downrange from the launch site. The test was previously slated to be accomplished with the Crew Dragon Demo-1 capsule;[498] but that test article exploded during a ground test of SuperDraco engines on 20 April 2019.[419] The abort test used the capsule originally intended for the first crewed flight.[499] As expected, the booster was destroyed by aerodynamic forces after the capsule aborted.[500] First flight of a Falcon 9 with only one functional stage — the second stage had a mass simulator in place of its engine.								
80	29 January 2020, 14:07[501]	F9 B5 B1051.3	CCAFS, SLC-40	Starlink 3 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]	LEO	SpaceX	Success	Success (drone ship)
	Third operational and fourth large batch of Starlink satellites, deployed in a circular 290 km (180 mi) orbit. One of the fairing halves was caught, while the other was fished out of the ocean.[502]								
81	17 February 2020, 15:05[503]	F9 B5 B1056.4	CCAFS, SLC-40	Starlink 4 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]	LEO	SpaceX	Success	Failure (drone ship)
	Fourth operational and fifth large batch of Starlink satellites. Used a new flight profile which deployed into a 212 km × 386 km (132 mi × 240 mi) elliptical orbit instead of launching into a circular orbit and firing the second stage engine twice. The first stage booster failed to land on the drone ship[504] due to incorrect wind data.[505] This was the first time a flight proven booster failed to land.								
82	7 March 2020, 04:50[506]	F9 B5 B1059.2	CCAFS, SLC-40	SpaceX CRS-20 (Dragon C112.3)	1,977 kg (4,359 lb)[507]	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
	Last launch of phase 1 of the CRS contract. Carries *Bartolomeo*, an ESA platform for hosting external payloads onto ISS.[508] Originally scheduled to launch on 2 March 2020, the launch date was pushed back due to a second stage engine failure. SpaceX decided to swap out the second stage instead of replacing the faulty part.[509] It was SpaceX's 50th successful landing of a first stage booster, the third flight of the Dragon C112 and the last launch of the cargo Dragon spacecraft.								
83	18 March 2020, 12:16[510]	F9 B5 B1048.5	KSC, LC-39A	Starlink 5 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]	LEO	SpaceX	Success	Failure (drone ship)
	Fifth operational launch of Starlink satellites. It was the first time a first stage booster flew for a fifth time and the second time the fairings were reused (Starlink flight in May 2019).[511] Towards the end of the first stage burn, the booster suffered premature shut down of an engine, the first of a Merlin 1D variant and first since the CRS-1 mission in October 2012. However, the payload still reached the targeted orbit.[512] This was the second Starlink launch booster landing failure in a row, later revealed to be caused by residual cleaning fluid trapped inside a sensor.[513]								
84	22 April 2020, 19:30[514]	F9 B5 B1051.4	KSC, LC-39A	Starlink 6 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]	LEO	SpaceX	Success	Success (drone ship)

Access the data collection scraped source file here:

https://github.com/crayoncannon/IBM-Data-Science-Professional-Certificate/blob/main/Final_Project/2.%20Data%20Collection%20%20Webscrape.ipynb

Data Wrangling

Steps taken:

1. **Calculate the number of launches on each site**
2. **Calculate the number and occurrence of each orbit**
3. **Calculate the number and occurence of mission outcome of the orbits**
4. **Create a landing outcome label from Outcome column**

Access the data wrangling source file here:

https://github.com/crayoncannon/IBM-Data-Science-Professional-Certificate/blob/main/Final_Project/3.%20Data%20Wrangling.ipynb

EDA with Data Visualization

Visualiztion between:

- Flight number and payload mass
- Flight number and launch site
- Payload mass and Launch site
- Orbit type
- Flight number and Orbit type
- Payload Mass and Orbit type

Access full visualization here:

https://github.com/crayoncannon/IBM-Data-Science-Professional-Certificate/blob/main/Final_Project/3.%20Data%20Wrangling.ipynb

EDA with Data Visualization

Example: Flight number vs payload mass:

1. As flight number increases, the first stage is more likely to land successfully
2. Even with more massive payloads, the first stage often returns successfully

EDA with SQL

Performed actions:

- **Find ames of the unique launch sites in the space mission**
- **Display 5 records where launch sites begin with the string 'CCA'**
- **Display the total payload mass carried by boosters launched by NASA (CRS)**
- **Display average payload mass carried by booster version F9 v1.1**
- **List the date when the first succesful landing outcome in ground pad was acheived.**
- **List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000**
- **List the total number of successful and failure mission outcomes**
- **List the names of the booster_versions which have carried the maximum payload mass. Use a subquery**

More performed actions can be found here:

https://github.com/crayoncannon/IBM-Data-Science-Professional-Certificate/blob/main/Final_Project/4.%20Exploratory%20Analysis%20SQL.ipynb

Build an Interactive Map with Folium

- Clustered launch site on a interactive map
- Each launch site is marked with the number of successful and unsuccessful launch.
- At a glance we can know how many past launches were carried at an area and whether they are successful or not
- Distance from launch site to Amenities and coastal area are measured to determine if the location of launch affects the success rate

Folium map can be found here:

https://github.com/crayoncannon/IBM-Data-Science-Professional-Certificate/blob/main/Final_Project/6.%20Data%20Visualization%20.ipynb

Build a Dashboard with Plotly Dash

- Pie chart shows success rate group according to launch area
- Scatter plot chart shows Success count on Payload mass for all sites according to booster version
- Dashboard easily shows variations

Dashboard code can be found here:

https://github.com/crayoncannon/IBM-Data-Science-Professional-Certificate/blob/main/Final_Project/7.%20Interactive%20dashboard%20Dash.py

Predictive Analysis (Classification)

Classification model used:

- Logistic regression
- Support Vector Machine
- Decision Tree Classifier
- K nearest Neighbor

Conclusion by comparing confusion matrix and accuracy scores:

Testing accuracy across all methods are almost the same except for decision tree method which performs the worst.

Analysis can be found here:

https://github.com/crayoncannon/IBM-Data-Science-Professional-Certificate/blob/main/Final_Project/8.%20Predictive%20Analysis.ipynb

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

- Show a scatter plot of Flight Number vs. Launch Site

Payload vs. Launch Site

- Show a scatter plot of Payload vs. Launch Site

Now if you observe Payload Mass Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

Success Rate vs. Orbit Type

- Show a bar chart for the success rate of each orbit type

Flight Number vs. Orbit Type

- Show a scatter point of Flight number vs. Orbit type

You can observe that in the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.

Payload vs. Orbit Type

- Show a scatter point of payload vs. orbit type

With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Launch Success Yearly Trend

- Show a line chart of yearly average success rate

you can observe that the sucess rate since 2013 kept increasing till 2020

All Launch Site Names

- Find the names of the unique launch sites

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (p
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (p
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	N
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	N
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	N

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

Sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

- Find the date of the first successful landing outcome on ground pad

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Booster_Version
F9 FT B1021.1
F9 FT B1022
F9 FT B1023.1
F9 FT B1026
F9 FT B1029.1
F9 FT B1021.2
F9 FT B1029.2
F9 FT B1036.1
F9 FT B1038.1
F9 B4 B1041.1
F9 FT B1031.2
F9 B4 B1042.1
F9 B4 B1045.1
F9 B5 B1046.1

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

Mission_Outcome	Total Number
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

Month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

Landing_Outcome	Count(Landing_Outcome)
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

<Folium Map Screenshot 1>

- Location of launch sites, easy to identify than having just the lattitudes

<Folium Map Screenshot 2>

- Location of launch sites, easy to identify the number of launch in a cluster

<Folium Map Screenshot 3>

- Location of launch sites, easy to identify the distance from amenities e.g. the coast

Section 4

Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>

- Success rate of a landing the first stage grouped by launch sites

- KSC LC-39A has the most number of successful landing

<Dashboard Screenshot 2>

- Success rate of KSC LC-39A

- KSC LC-39A has the highest launch success ratio

<Dashboard Screenshot 3>

- Success rate of payload mass on all sites

- Ft gives more consistent result across all payload

Section
5

Predictive Analysis (Classification)

Classification Accuracy

- Visualize the built model accuracy for all built classification models
- Find which model has the highest classification accuracy

Confusion Matrix

- Logistic regression, SVM KNN all has similar accuracy on the confusion matrix the best performing model with an explanation

- True Postive - 12 (True label is landed, Predicted label is also landed)
- False Postive - 3 (True label is not landed, Predicted label is landed)

Conclusions

- Testing accuracy across all methods are almost the same except for decision tree method which perfoms the worst.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

- Github link to all resources:
- https://github.com/crayoncannon/IBM-Data-Science-Professional-Certificate/tree/main/Final_Project

Thank you!