Winning Space Race with Data Science

Rogério Yamada

22 September 2025

IBM Developer
SKILLS NETWORK

Outline

Executive Summary

Introduction

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology
  - Data about rocket launches was obtained from a SpaceX API and web scraping Wikipedia pages
- Perform data wrangling
  - Missing data was handled, a preliminary Exploratory Data Analysis was performed, and the variable `Outcome Class` was defined for training the supervised models
- Perform Exploratory Data Analysis (EDA) using visualisation and SQL
- Perform interactive visual analysis using Folium and Plotly Dash
- Perform predictive analysis using classification models

Data Collection

- SpaceX API data extraction and Wikipedia pages web scraping were combined to produce a dataset of SpaceX Falcon 9 landings information

data_falcon9

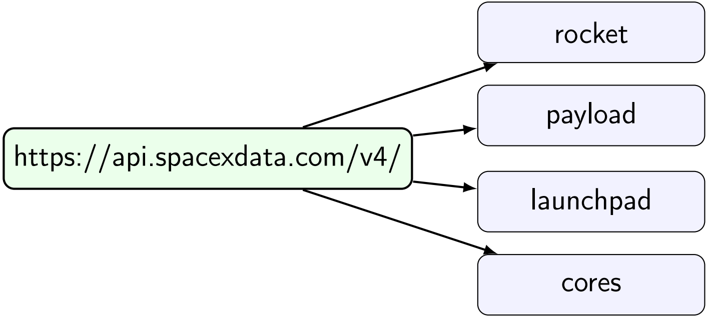
[34]

11 rows 90 rows × 17 cols

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
0	1	2010-06-04	Falcon 9	6123.547647	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857
1	2	2012-05-22	Falcon 9	525.000000	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005	-80.577366	28.561857
2	3	2013-03-01	Falcon 9	677.000000	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007	-80.577366	28.561857
3	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003	-120.610829	34.632093
4	5	2013-12-03	Falcon 9	3170.000000	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004	-80.577366	28.561857
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
85	86	2020-09-03	Falcon 9	15600.000000	VLEO	KSC LC 39A	True ASDS	2	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	12	B1060	-80.603956	28.608058
86	87	2020-10-06	Falcon 9	15600.000000	VLEO	KSC LC 39A	True ASDS	3	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	13	B1058	-80.603956	28.608058
87	88	2020-10-18	Falcon 9	15600.000000	VLEO	KSC LC 39A	True ASDS	6	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	12	B1051	-80.603956	28.608058
88	89	2020-10-24	Falcon 9	15600.000000	VLEO	CCSFS SLC 40	True ASDS	3	True	True	True	5e9e3033383ecbb9e534e7cc	5.0	12	B1060	-80.577366	28.561857
89	90	2020-11-05	Falcon 9	3681.000000	MEO	CCSFS SLC 40	True ASDS	1	True	False	True	5e9e3032383ecb6bb234e7ca	5.0	8	B1062	-80.577366	28.561857

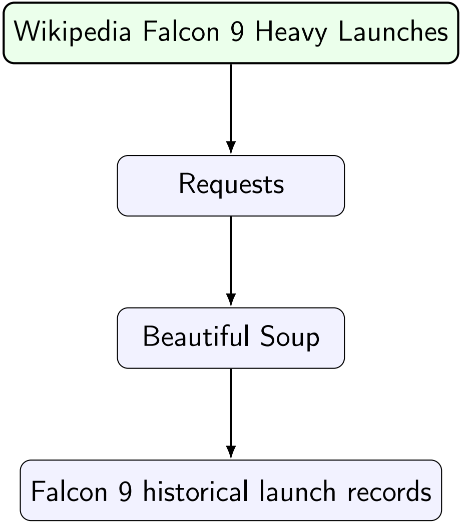
Data Collection – SpaceX API

- From the SpaceX API endpoint `https://api.spacexdata.com/v4/` we probed the following data sources:
    - rocket
    - payload
    - launchpad
    - cores
- Jupyter Notebook's GitHub URL



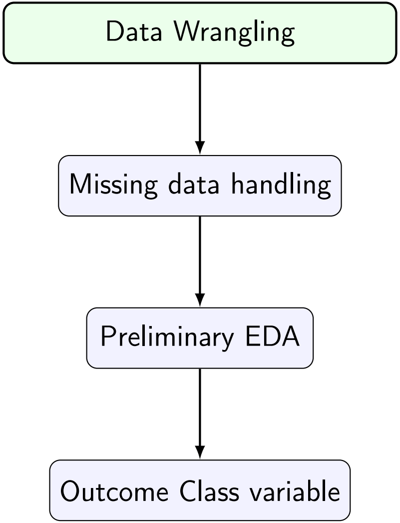
Data Collection – Scraping

- From the Wikipedia List of Falcon 9 and Falcon Heavy launches web page we collected Falcon 9 historical launch records
- Jupyter Notebook's GitHub URL



Data Wrangling

- Through data wrangling, the variable `Outcome Class` was defined for training the supervised models
- Jupyter Notebook's GitHub URL



EDA with SQL

- Several SQL queries have been processed to gain insights about the landing outcomes:
  - Launching sites
  - Total payload mass carried by specific boosters in specific sites
  - Successful and failed landing outcomes
- Jupyter Notebook's GitHub URL

EDA with Data Visualisation

- Several charts have been drawn to gain insights about the landing outcomes:
  - Flight Number versus Launch Site by Class
  - Payload Mass versus Launch Site by Class
  - Success Rate by Orbit
  - Flight Number versus Orbit by Class
  - Payload Mass versus Orbit by Class
  - Yearly Launch Success Rate
- Jupyter Notebook's GitHub URL

Build an Interactive Map with Folium

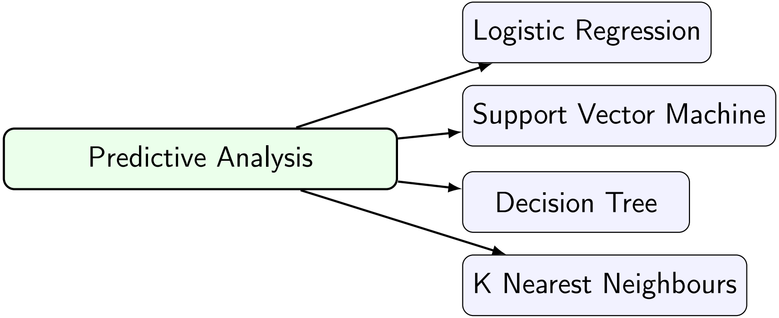
- A geographical analysis has been performed to gain insights about the dependencies of landing outcomes and location and surrounding of launching sites:
  - Marking all launch sites on a map
  - Marking the success/failed launches for each site on the map
  - Calculating the distances between a launch site to its proximities
- Jupyter Notebook's GitHub URL

Build a Dashboard with Plotly Dash

- A dashboard has been implemented to perform real-time analysis about landing outcomes considering:
  - Launch site drop-down menu
  - Interactive successful landing outcome pie chart
  - Range slider for selecting payload mass
  - Interactive successful landing outcome scatter plot
- Jupyter Notebook's GitHub URL

Predictive Analysis (Classification)

- The following predictive models have been considered:
  - Logistics Regression
  - Support Vector Machine
  - Decision Tree
  - K Nearest Neighbours
- Jupyter Notebook's GitHub URL



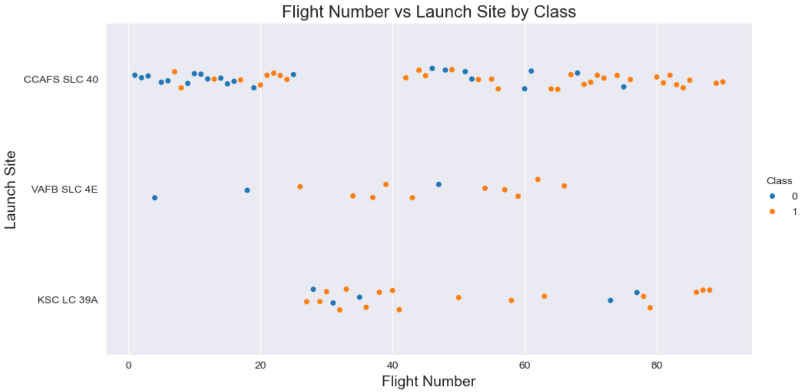
Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

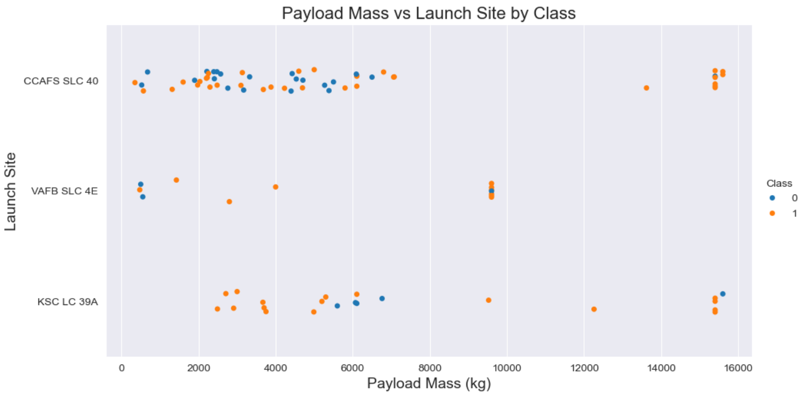
Insights drawn from EDA

Flight Number versus Launch Site



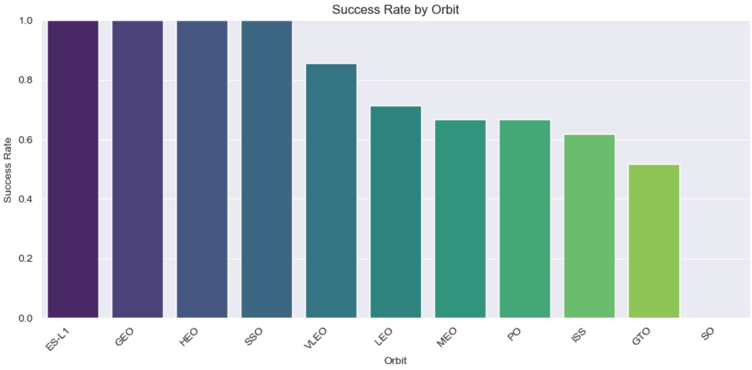
- There is a concentration of unsuccessful landing outcomes in the site CCAFS SLC 40
- As the number of flights increase, the successful landing outcomes are likely to increase

Payload versus Launch Site



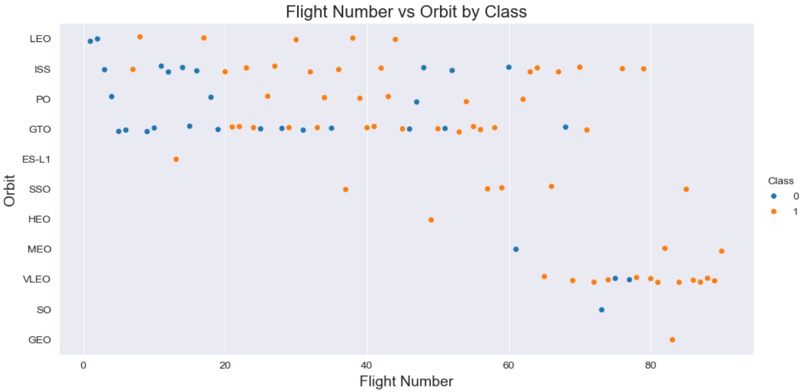
- There is a concentration of unsuccessful landing outcomes in CCAFS SLC 40 for payloads lighter than 7000 kg

Success Rate versus Orbit Type



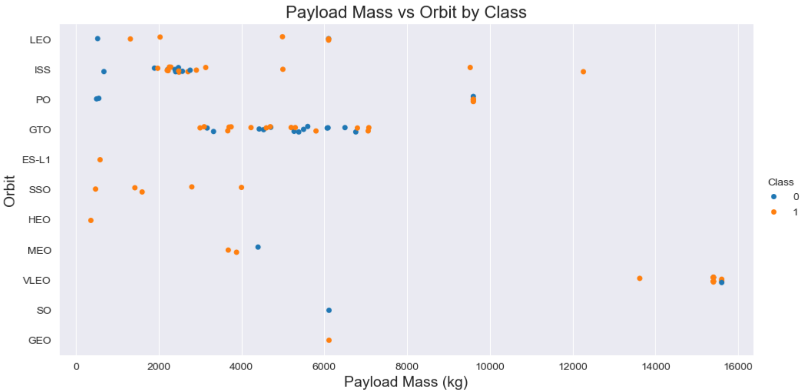
- The success rate is higher when the orbit type is ES-L1, GEO, HEO, and SSO
- The success rate is lower when the orbit type is GTO and SO

Flight Number versus Orbit Type



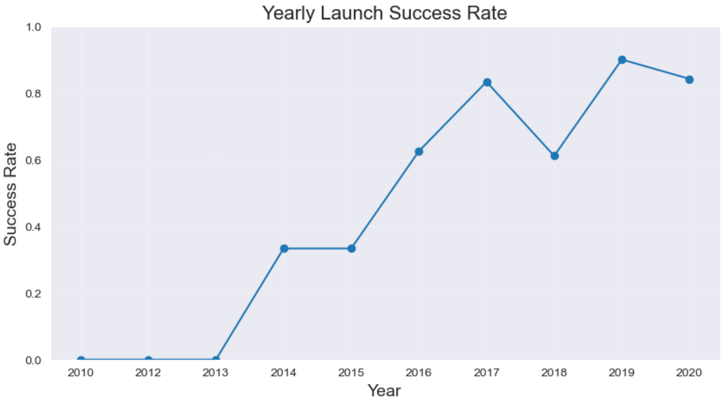
- The success rate is higher as the number of flights increases for orbit type LEO, ISS, and PO
- There seems to be no relationship between flight number when the orbit type is GTO

Payload versus Orbit Type



- With heavy payloads, the successful landing rate is higher for PO, LEO, and ISS orbit types
- However, regarding GTO, we cannot distinguish this well as both positive landing rate and negative landing (unsuccessful mission) occur almost evenly

Launch Success Yearly Trend



- The success rate since 2013 kept increasing until 2017 (stable in 2014) and after 2015 it started increasing

All Launch Site Names

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

- This is a sorted list of unique, non-null values from the Launch_Site column in the SPACEXTABLE table

Launch Site Names Begin with ‘CCA’

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
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 (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- This is a set of up to 5 rows from the SPACEXTABLE table where the Launch_Site column starts with the text “CCA” (e.g., “CCAFS ...”). The LIKE ‘CCA%’ filter matches any value beginning with “CCA”

Total Payload Mass

Total Payload Mass = 45596

- This is the total payload mass (sum of PAYLOAD_MASS__KG_) for all rows in SPACEXTABLE where the Customer is exactly 'NASA (CRS)'. The result is a single value labeled total_payload_mass

Average Payload Mass by F9 v1.1

Average Payload Mass = 2534.6666666666665

- This is the average payload mass (PAYLOAD_MASS__KG_) over rows in SPACEXTABLE whose Booster_Version starts with “F9 v1.1”, returning a single value named avg_payload_mass

First Successful Ground Landing Date

First Ground Pad Landing Success Date = 2015-12-22

- This is the earliest Date (minimum) among rows where Landing_Outcome is exactly “Success (ground pad)”, labeled first_ground_pad_success_date.

Successful Drone Ship Landing with Payload between 4000 and 60000

booster_name
F9 FT B1021.2
F9 FT B1031.2
F9 FT B1022
F9 FT B1026

- This are the unique booster versions (as booster_name) that had a “Success (drone ship)” landing and carried payloads strictly between 4000 kg and 6000 kg, sorted alphabetically

Total Number of Successful and Failure Mission Outcomes

Boosters Carried Maximum Payload

2015 Launch Records

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

Section 4

Launch Sites Proximities Analysis

<Folium Map Screenshot 1>

<Folium Map Screenshot 2>

Folium Map Screenshot 3

Section 5

Build a Dashboard with Plotly Dash

Dashboard Screenshot 1

Dashboard Screenshot 2

Dashboard Screenshot 3

Section 6

Predictive Analysis (Classification)

Classification Accuracy

Confusion Matrix

Conclusions

-

Appendix

Thank you!