IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Samuel Pariente
2021/10/26

Outline

Executive Summary

Methodology

- Data was collected through a combination of calls to the public SpaceX API and web scraping on Wikipedia
- Missing values in the collected data were replaced by observed averages, and landing successes and failures were clearly identified
- **Exploratory data analysis** (EDA) was performed, using visualization and SQL, as well as interactive visual analytics using Folium and Plotly Dash
- A **predictive analysis** using 4 separate classification models was performed

Results

- Among the insights drawn from the EDA, **possible relationships were identified** between successful landings and payload site, payload weight, target orbit, and launch date
- The resulting predictive classification models were able to **predict the landing outcome with ca. 83% accuracy** on a test set, and can accurately be used when predicting a failure to land

Github repository: https://github.com/SPariente/IBM_DS_capstone

Introduction

Situation overview

- SpaceX advertises Falcon 9 rocket launches with a cost of 62 million dollars, where other providers cost upward of 165 million dollars each
- Much of the savings is linked to SpaceX's ability to reuse the first stage of the rocket

Objectives

- By predicting if the first stage will land, it is possible to determine the cost of a launch
- This information can be used for competition purposes by an alternate company, to decide whether to bid against SpaceX for a rocket launch
- To this end, we aim to assess whether the landing outcome of SpaceX launches can be determined using other launch features, and ultimately whether a model can be created to accurately predict this outcome for future launches

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology
  - Data was collected through a combination of calls to the public SpaceX API and web scraping on Wikipedia
- Data wrangling approach
  - Missing values in the collected data were replaced by observed averages, and landing successes and failures were clearly identified
- Exploratory data analysis (EDA)
  - Data visualization was used for exploratory data analysis purposes, to look at possible predictors of landing success
- Perform interactive visual analytics using Folium and Plotly Dash
  - Building on the EDA, an interactive dashboard was created, to dive deeper into the data using selective criteria
- Predictive analysis using classification models
  - To predict the outcome of a launch, 4 different machine learning models were trained, using separate training and test sets to assess accuracy

Data Collection

Data on all launches using the Falcon 9 booster was obtained from publicly available online sources

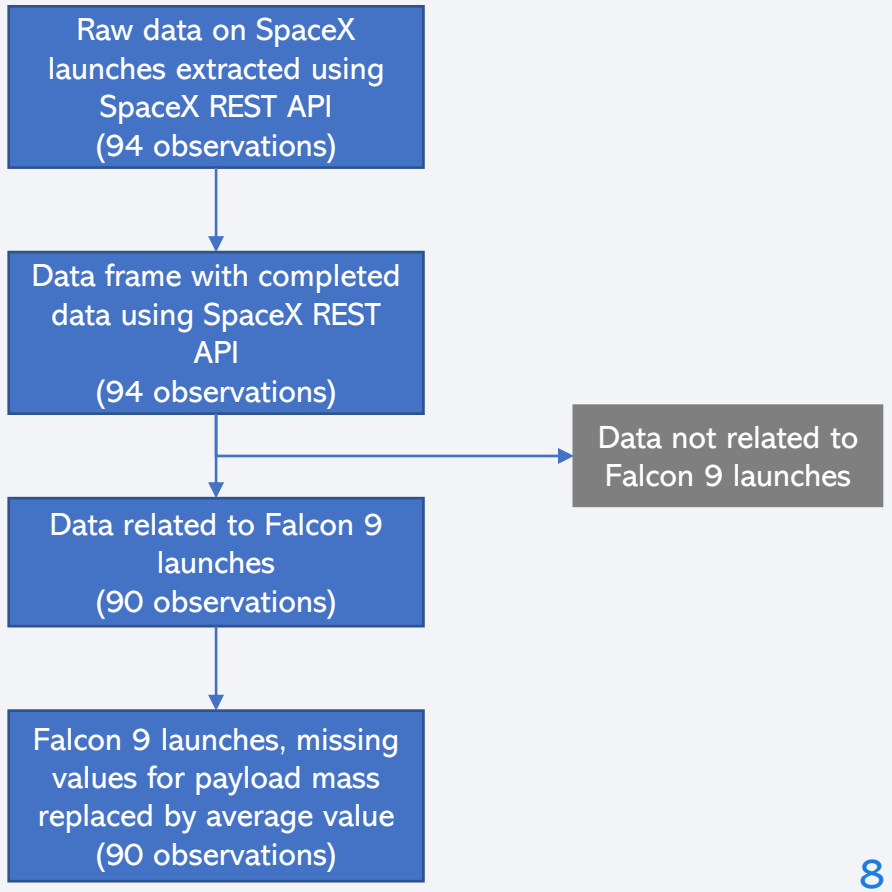
- Calls to the SpaceX REST API were used to obtain detailed information on 90 Falcon 9 launches, including technical specifications, launch date and site, target orbit, and payload mass
- Web scraping was used on the “List of Falcon 9 and Falcon Heavy launches” Wikipedia page to complete the data, with additional elements including the booster version and the customer name

Data Collection – SpaceX API

Methodology

- Using Python, successive calls were made to the public SpaceX REST API to extract:
  - A list of SpaceX launches
  - The names of the rockets used
  - Additional data on the launchpad used
  - Additional data on the payload launched
  - Additional data on the rocket cores
- The resulting dataset was converted into a data frame for further wrangling and use
- Data related to Falcon 9 launches was kept, with other observations (Falcon 1) removed from the dataset
- 5 missing values for payload mass were replaced by the average payload mass across the sample

Flowchart overview



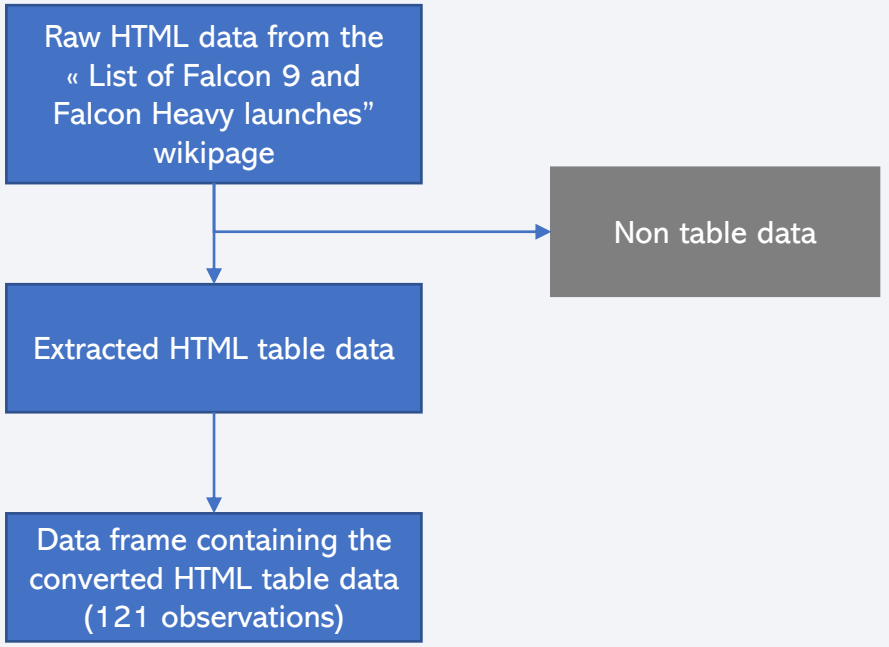
Github repository: https://github.com/SPariente/IBM_DS_capstone/blob/master/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection – Scraping

Methodology

- Using Python, the Wikipedia page titled “List of Falcon 9 and Falcon Heavy launches” was scraped
  - A table containing Falcon 9 records was extracted, after identification of the relevant headers
  - The table was then converted into a data frame for wrangling and future use

Flowchart overview



Github repository: https://github.com/SPariente/IBM_DS_capstone/blob/master/jupyter-labs-webscraping.ipynb

Data Wrangling

Insights from preliminary exploratory data analysis

- 3 distinct launch sites
- 11 target orbits
- 8 different landing outcomes

Data wrangling approach using Python

- Landing outcomes values split between successful landings and failed landings
- New binary field added to the data frame, with 1 assigned to successful landings and 0 to failed landings

Github repository: https://github.com/SPariente/IBM_DS_capstone/blob/master/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Data visualization was used for exploratory data analysis purposes
- Several scatter plots were created, to visualize possible relationships:
  - Flight number by launch site, showing that a single site was used initially, and that 2 sites have a higher success rate for landing than the third
  - Payload by launch site, showing that the three sites tend to be used for different types of payload
  - Success rate by orbit type, showing that some orbits tend to have a higher success rate than others
  - Flight number by orbit type, showing that progressively more orbits have been used for the Falcon 9 missions, with an increasing success rate
  - Payload by orbit type, highlighting specific payload ranges for each target orbit
- A line chart showing the overall landing success rate over time was also created, highlighting a progressive increase in success rate since 2013

Github repository: https://github.com/SPariente/IBM_DS_capstone/blob/master/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- Using SQL, a dataset containing a record for each payload carried during a SpaceX mission into outer space was explored, with the following highlights:
  - 4 unique launch sites were identified
  - The records for 5 launches from launch sites the name of which started by 'CCA' were extracted
  - The total weight carried by NASA boosters was computed
  - The average payload mass carried by booster version F9 v1.1 was computed
  - The date of the first successful landing on a ground pad was extracted
  - The names of all boosters that have successfully landed on a drone ship were extracted
  - The total number of successful and failed missions was extracted
  - The names of each booster version that has carried the maximum payload mass were extracted
  - For every month in 2017 that saw such an occurrence, the booster version and launch site of each mission that had a successful landing on a ground pad was extracted
  - The total amount of successful landings between 4th June 2010 and 20th March 2017 were ranked by outcome, in descending order

Github repository: https://github.com/SPariente/IBM_DS_capstone/blob/master/jupyter-labs-eda-sql-edx-sanitized.ipynb

Build an Interactive Map with Folium

- Based on a dataset that included latitude and longitude information of the launch site for each Falcon 9 mission, interactive maps were created using the Folium package in Python
- Each map was built to highlight a specific element:
  - A map showing the location of the 4 launch sites, with a circle and a marker for each
  - A map with added marker clusters, showing the landing outcome (success of failure) of every launch for each site, to analyze whether geographical location of the launch could have had an impact on the success rate
  - A map, with added markers and lines showing the distance, for a specific site, to the closest railway and highway, to assess whether this can have an impact on the choice of location

Github repository: https://github.com/SPariente/IBM_DS_capstone/blob/master/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- To facilitate the access to several insights from the exploratory data analysis, a dashboard was created in Python
- The dashboard shows 2 figures:
  - A pie chart showing either the distribution of successful launches across all sites, or the proportion of successful launches for a specific site
  - A scatter plot showing the distribution of successful launches across a range of payload masses, for a specific site if selected, color coded by booster version
- The user can select to view either all launch sites or a specific one, through a dropdown list, to get detailed information on a specific site
- For the payload scatter plot, the user can further select a specific payload mass range through a slider, to gain a deeper understanding of specific mass ranges

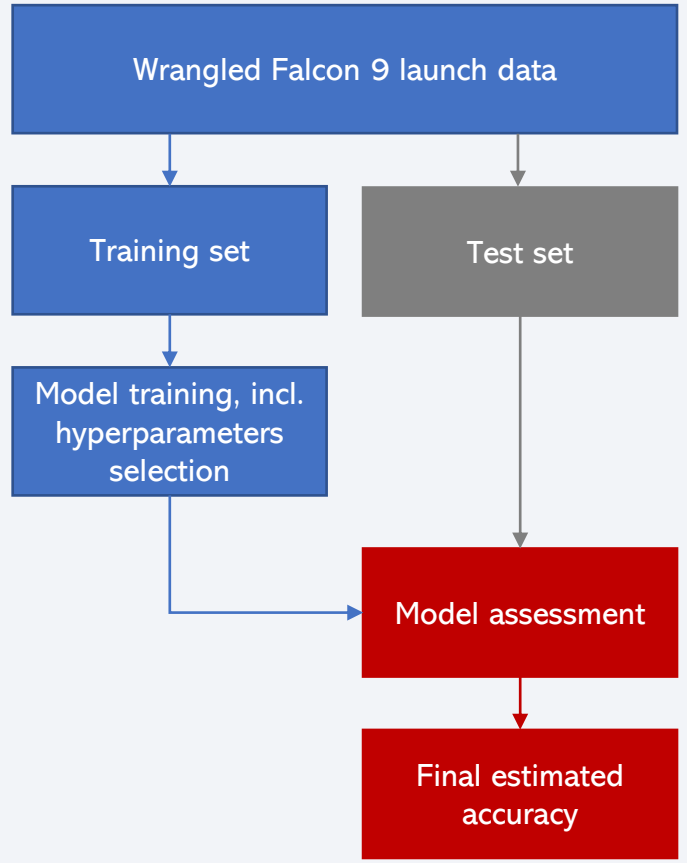
Github repository: https://github.com/SPariente/IBM_DS_capstone/blob/master/spacex_dash_app_final.py

Predictive Analysis (Classification)

Methodology

- In order to predict the landing outcome of a launch, 4 different machine learning models were trained
- The training process involved:
  - Splitting the available data into training and test sets
  - Finding the best hyperparameters for each model
  - Training the machine learning algorithms
- After training using the best hyperparameters, all 4 models resulted in the same accuracy of ca. 83% on the test set.
- No combination of these 4 models could improve this accuracy, as they all yield the same exact predictions.

Flowchart overview

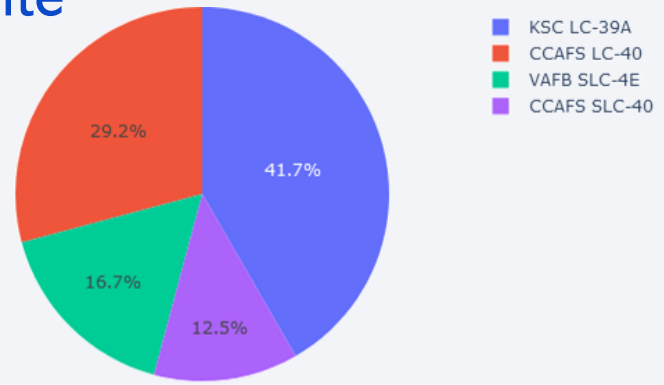


Github repository: https://github.com/SPariente/IBM_DS_capstone/blob/master/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

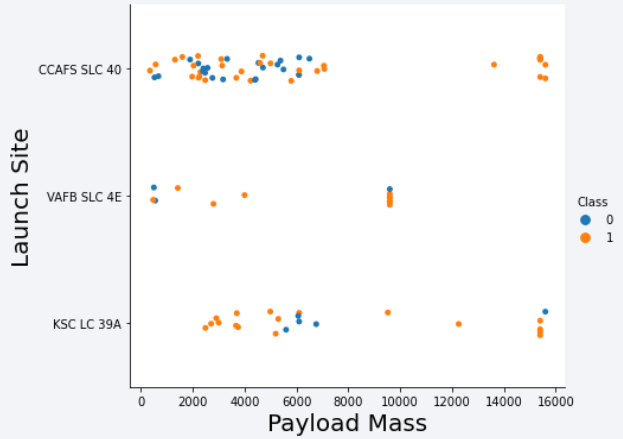
Results

- The exploratory data analysis showed that success rate is influenced by:
  - The launch date, with increasing success rate with time
  - The launch site, with some sites more successful than others
  - The target orbit, with some orbit materially more successful than others
- The payload weight, to a lesser extent, but especially when coupled with the target orbit
- After training 4 different machine learning models, each was able to accurately predict ca. 83% of the landing outcomes on a test set.

Share of successful landings by launch site

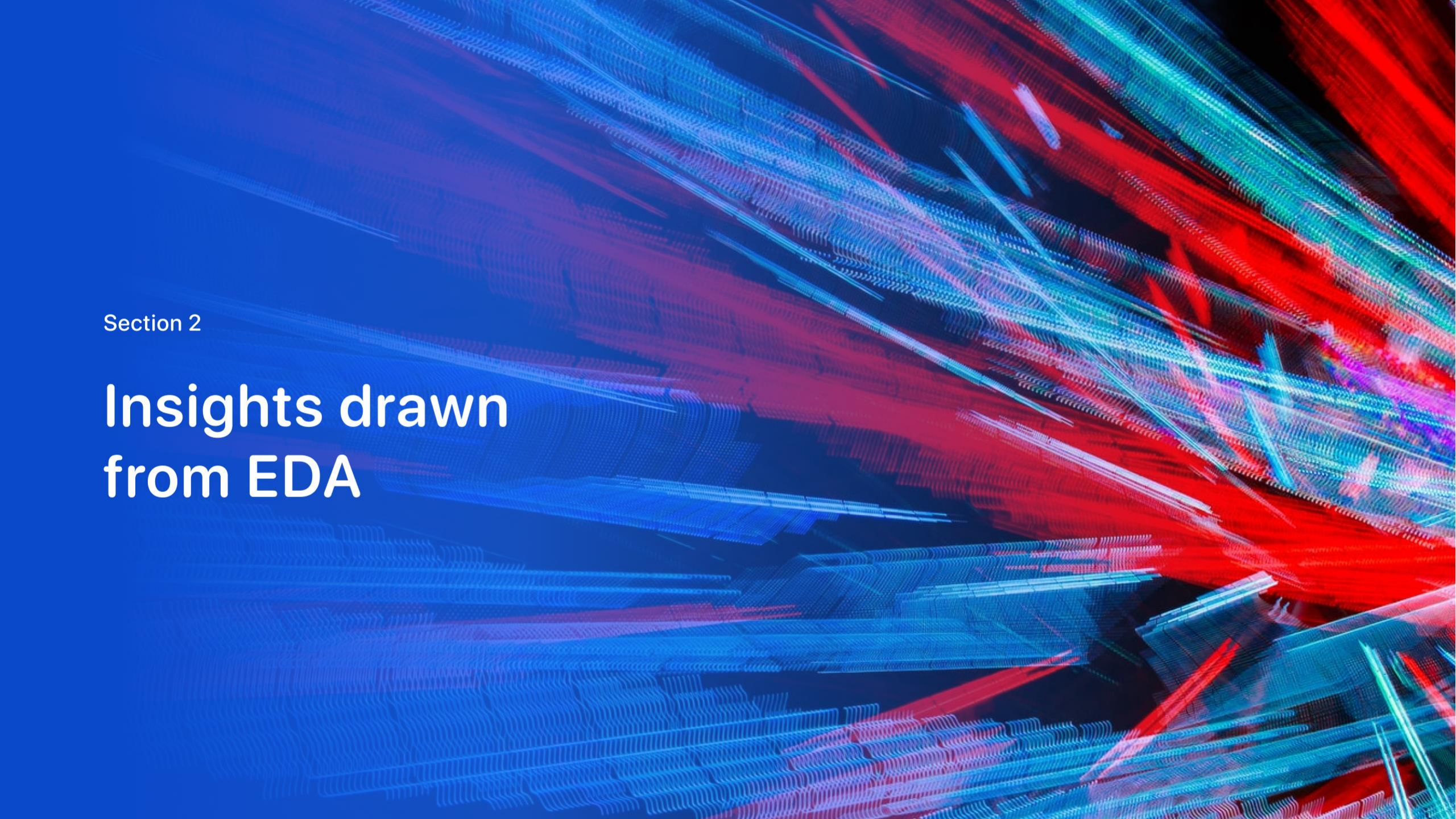


Payload mass impact on landing



Section 2

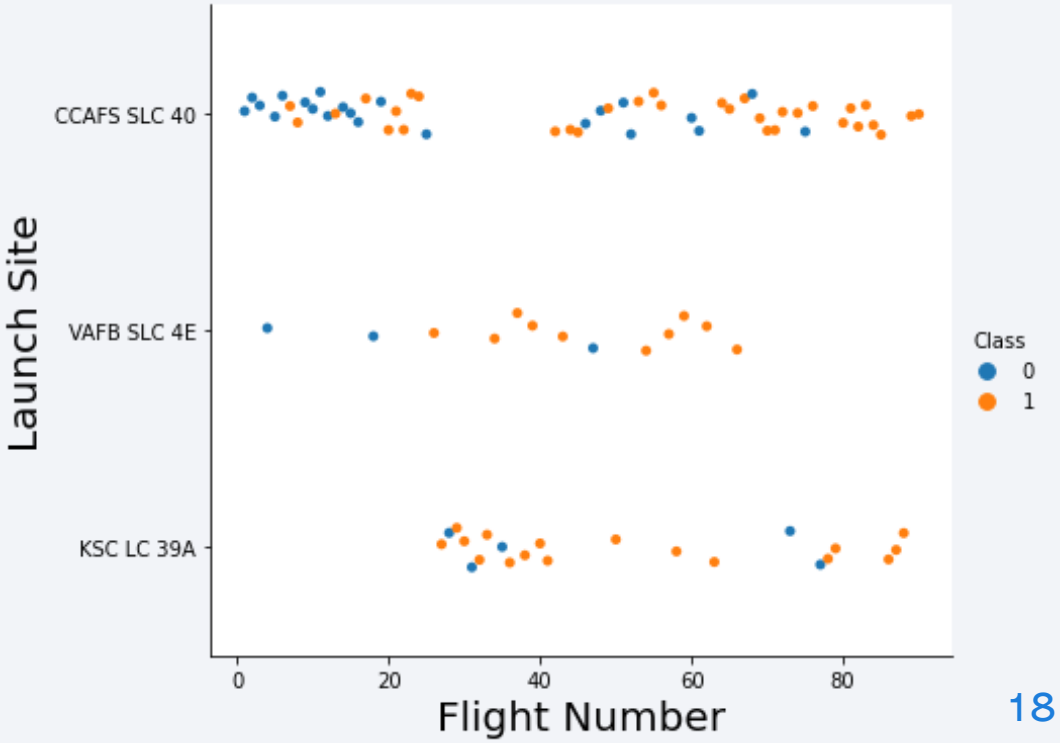
Insights drawn from EDA



Flight Number vs. Launch Site

- Initially, almost all flights were made using the launch site CCAFS SLC 40
- Site VAFB SLC 4E seems to be less used than the others, especially for recent flights
- More recent flights tend to have a higher success rate than older ones
- Flights launched from sites VAFB SLC 4E and KSC LC 39A have the highest success rates at ca. 77%, vs. 60% for site CCAFS SLC 40

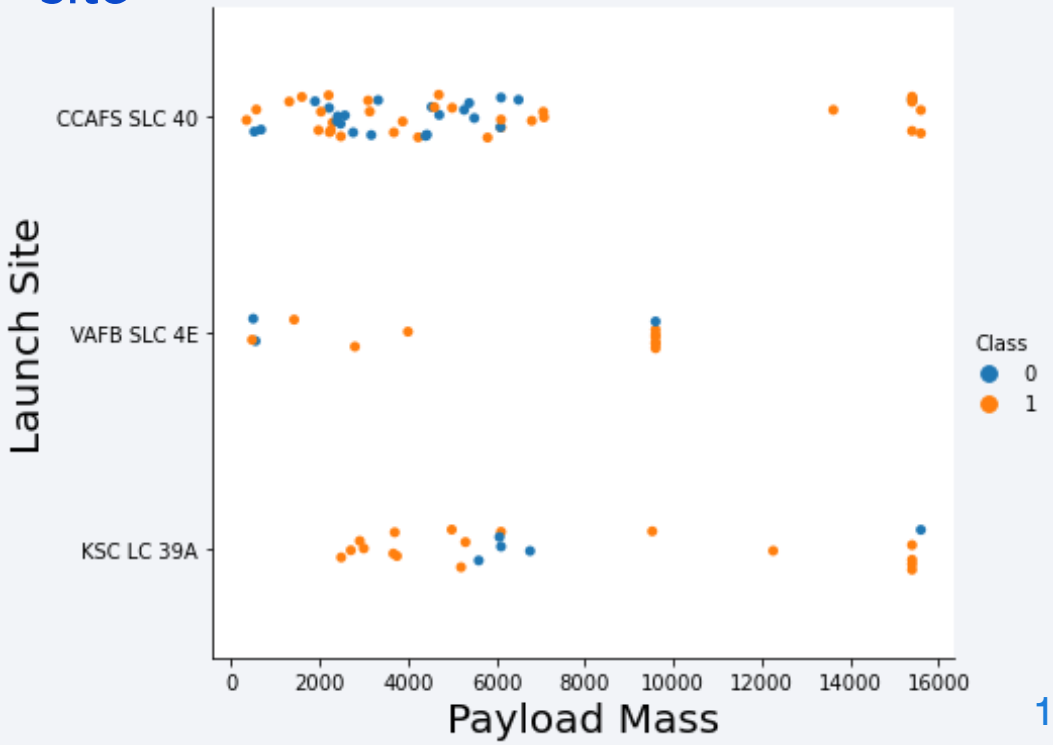
Overview of successful (1) and failed (0) landings by flight number and launch site



Payload vs. Launch Site

- Flights with a payload weighing less than 2,000 kg are made from launch sites CCAFS SLC 40 and VAFB SLC 4E
- Flights with a payload weighing more than 10,000 kg are made from launch sites CCAFS SLC 40 and KSC LC 39A
- Launches with a payload weighing more than 8,000 kg tend to have a higher success rate (87%) than those with a lighter payload (60%), especially if launched from CCAFS SLC 40 (100% success rate above 8,000 kg)

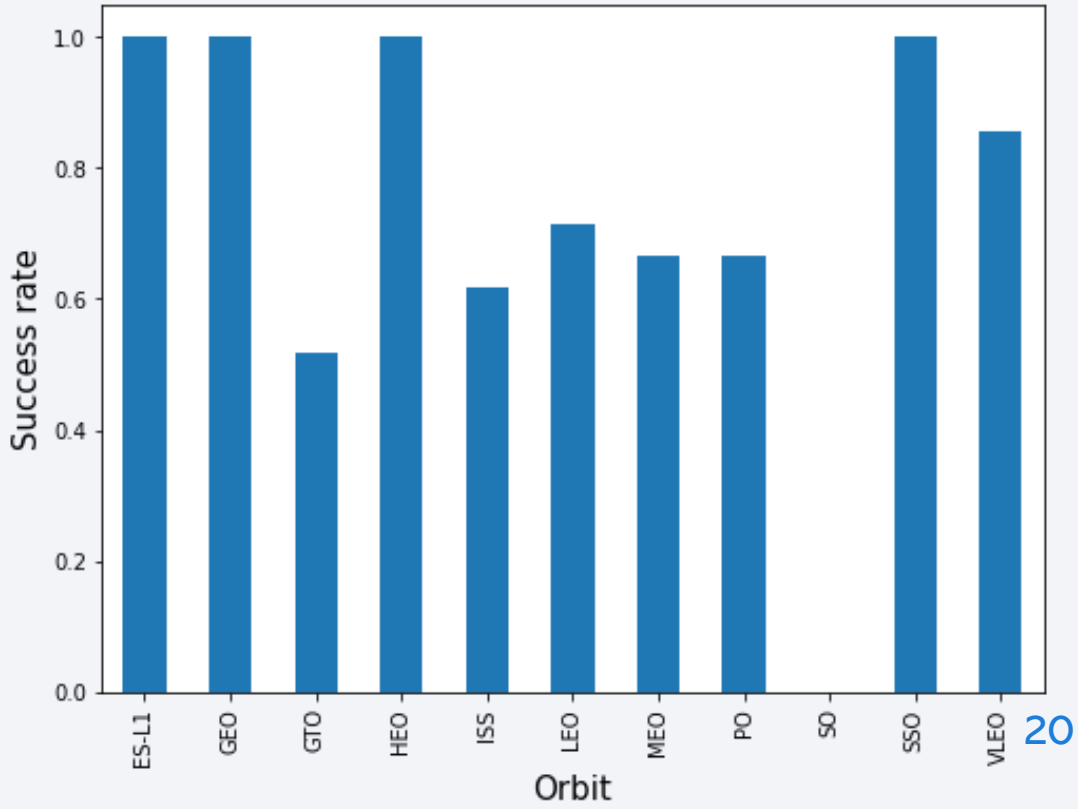
Overview of successful (1) and failed (0) landings by payload mass and launch site



Success Rate vs. Orbit Type

- Flights with a target orbit being ES-L1 (L1 Lagrange point), GEO (geosynchronous), HEO (highly elliptical), or SSO (helio-synchronous) have the higher success rate.
- It should be noted that out of these, only SSO has had more than 1 launch, and that, by including SO (also heliosynchronous), its success rate falls to 83%, behind VLEO (very low Earth orbit) at 86%.
- Flights to GTO (geosynchronous) and to the ISS have the lowest success rates, at 52% and 62% respectively, but also the highest number of flights overall at 27 and 21, respectively.

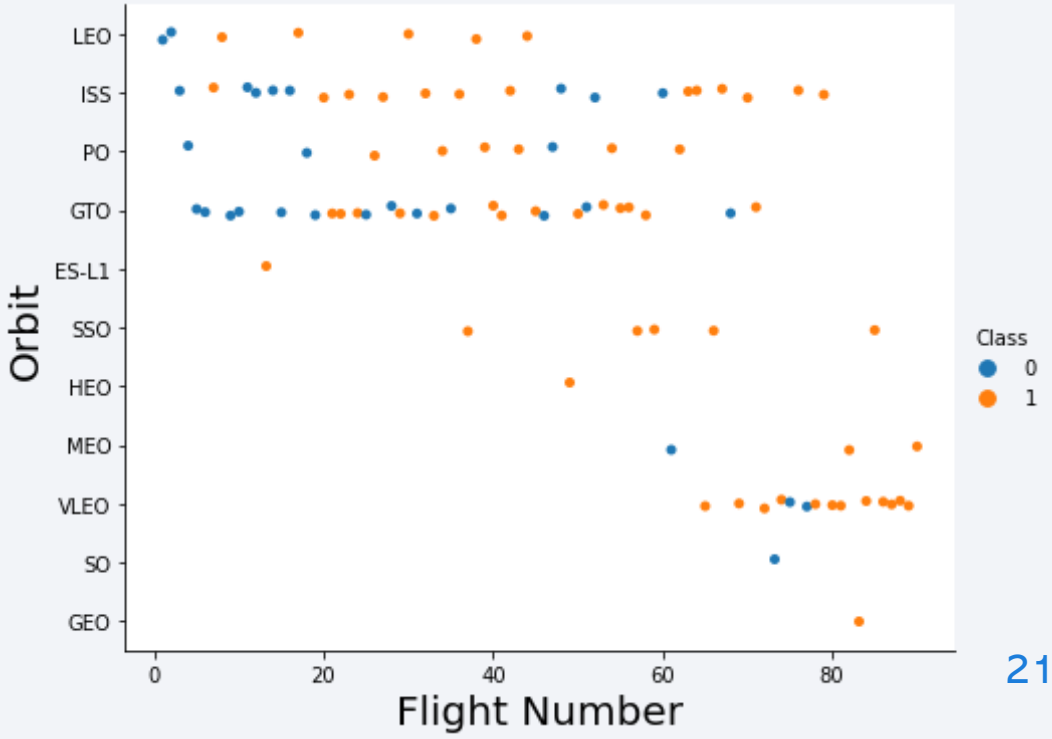
Ratio of successful landings by target orbit



Flight Number vs. Orbit Type

- Initial flights were focused on the LEO (low Earth orbit), ISS, PO (polar), and GTO orbits
- Recent flights have notably included VLEO, SSO, and MEO (intermediate circular orbit), with relatively high success rates compared to older flights
- The first launch to a new orbit tends to result in a failure to land, with a 55% failure rate

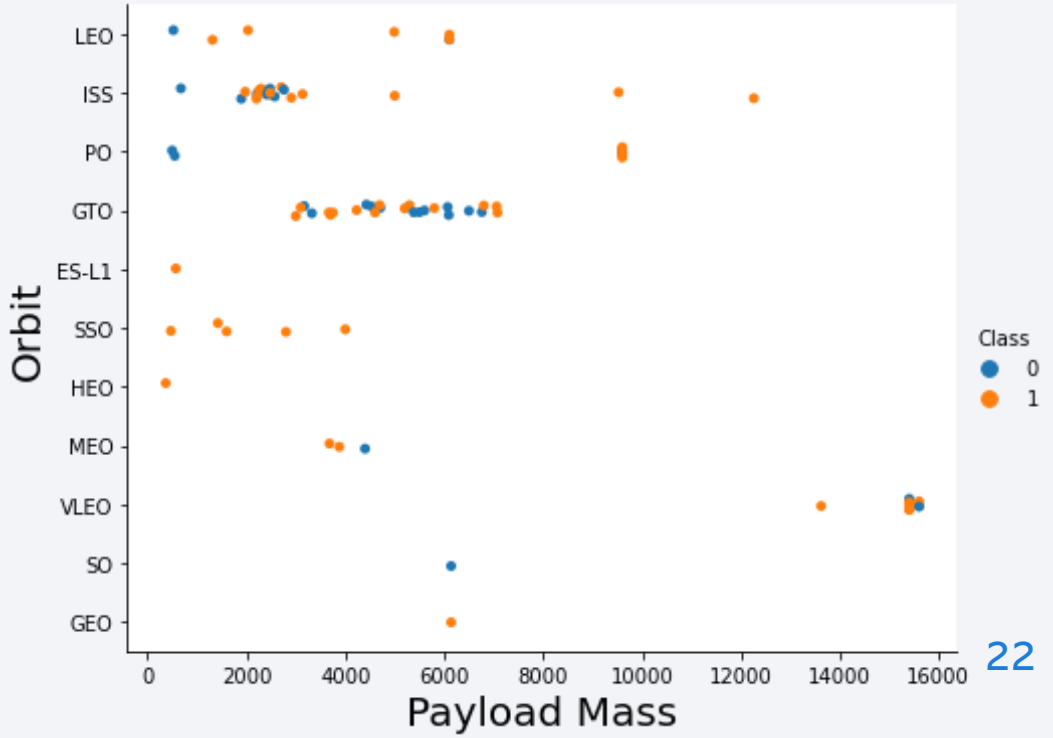
Overview of successful (1) and failed (0) landings by flight number and target orbit



Payload vs. Orbit Type

- Flights with payloads weighing less than 8,000 kg are used for most orbits, except VLEO
- Flights with payloads heavier than 8,000 kg are used for ISS, PO and VLEO orbits exclusively, and tend to have a higher success rate
- For GTO orbit launches, heavier payloads tend to have a lower success rate than lighter ones

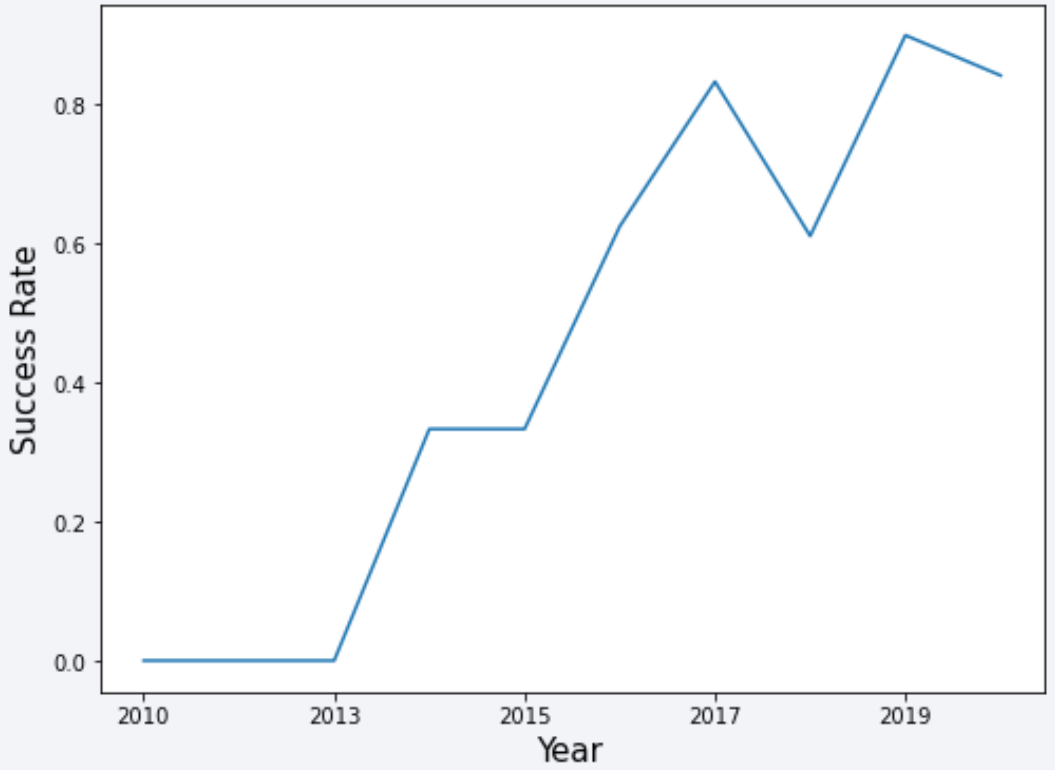
Overview of successful (1) and failed (0) landings by payload mass and target orbit



Launch Success Yearly Trend

- While initially mostly unsuccessful in the recovery process, the success rate of launches has steadily increased over the 2013-2017 period, to stabilize around 78% between 2017 and 2020

Overview of success rate over time



All Launch Site Names

Methodology

- A dataset containing a record for each payload carried during a SpaceX mission into outer space was used to obtain the names of all launch sites used

4 launch sites were identified

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

Launch Site Names Begin with 'CCA'

Methodology

- A dataset containing a record for each payload carried during a SpaceX mission into outer space was used to obtain information of 5 records where launch sites begin with `CCA`

Excerpt from the resulting table

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

Total Payload Mass

Methodology

- A dataset containing a record for each payload carried during a SpaceX mission into outer space was used to obtain the total payload carried by boosters from NASA

Total weight carried

- A total of 48,213 kg was carried by boosters from NASA

Average Payload Mass by F9 v1.1

Methodology

- A dataset containing a record for each payload carried during a SpaceX mission into outer space was used to obtain the average payload mass carried by booster version F9 v1.1

Average payload mass carried

- Booster version F9 v1.1 carried on average 2,928 kg per launch

First Successful Ground Landing Date

Methodology

- A dataset containing a record for each payload carried during a SpaceX mission into outer space was used to obtain the date of the first successful landing outcome on ground pad

First successful ground pad landing

- On the 22nd December 2015, the first successful landing on a ground pad by a SpaceX mission was achieved

Successful Drone Ship Landing with Payload between 4,000 and 6,000 kg

Methodology

- A dataset containing a record for each payload carried during a SpaceX mission into outer space was used to obtain the names of boosters which have successfully landed on drone ship and had a payload mass greater than 4,000 kg but less than 6,000 kg

4 separate boosters have had successful landings in this range

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

Total Number of Successful and Failure Mission Outcomes

Methodology

- A dataset containing a record for each payload carried during a SpaceX mission into outer space was used to calculate the total number of successful and failure mission outcomes

Overview of mission success

- Out of 101 missions :
  - 1 resulted in a failure in flight
  - 99 were a success
  - 1 was a success but with an unclear payload status

Boosters Carried Maximum Payload

Methodology

- A dataset containing a record for each payload carried during a SpaceX mission into outer space was used to obtain the names of the boosters which have carried the maximum payload mass

12 boosters have carried the maximum reported payload mass

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

2017 Launch Records

Methodology

- A dataset containing a record for each payload carried during a SpaceX mission into outer space was used to obtain a list of records for the months in 2017 that had a successful landing outcome on a ground pad

Overview of the 2017 successful ground pad landings by month

MONTH	landing__outcome	booster_version	launch_site
February	Success (ground pad)	F9 FT B1031.1	KSC LC-39A
May	Success (ground pad)	F9 FT B1032.1	KSC LC-39A
June	Success (ground pad)	F9 FT B1035.1	KSC LC-39A
August	Success (ground pad)	F9 B4 B1039.1	KSC LC-39A
September	Success (ground pad)	F9 B4 B1040.1	KSC LC-39A
December	Success (ground pad)	F9 FT B1035.2	CCAFS SLC-40

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

Methodology

- A dataset containing a record for each payload carried during a SpaceX mission into outer space was used to obtain and rank the count of successful landing outcomes between the 4th June 2010 and 20th March 2017.

Ranked successful landing outcomes

- In descending order, the number of successful landings by outcome between these dates is as follows:
  - Successful landing on a drone ship: 5 landings
  - Successful landing on a ground pad: 3 landings

Section 4

Launch Sites Proximities Analysis

Falcon 9 rocket launch sites locations

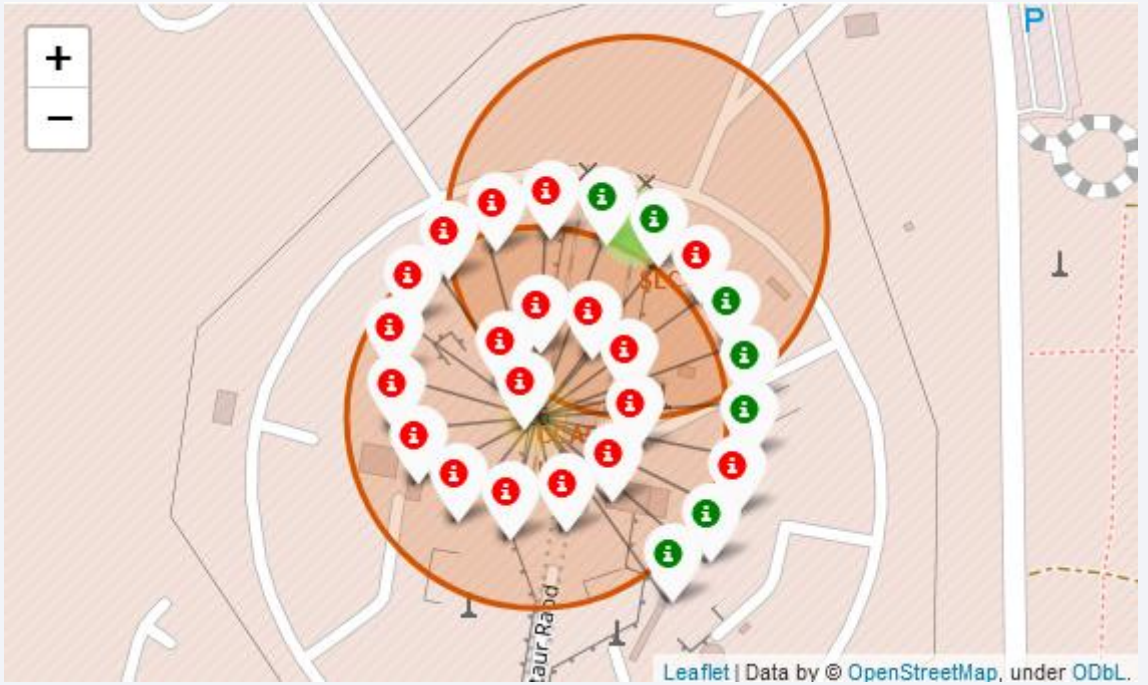
- The Falcon 9 rocket uses 4 launch sites, located in the continental US
  - 3 of them (CCAFS LC-40, CCAFS SLC-40, and KSC LC-39A) are located at the Cape Canaveral / Kennedy Space center facility, in Florida, at latitudes around 28.6° North
  - VAFB SLC-4E is located in the Vanderberg base in southern California, as a slightly higher latitude of 34.6° North

Mapped locations of Falcon 9 launches

Landing outcomes for flights out of CCAFS LC-40 launch site

- 52 flights were launched the CCAFS LC-40 launch site
- While most resulted in a failed or non-recoverable landing, a trend can be seen where the more recent launches (external end of the spiral) have a higher success rate than earlier ones (center of the spiral)

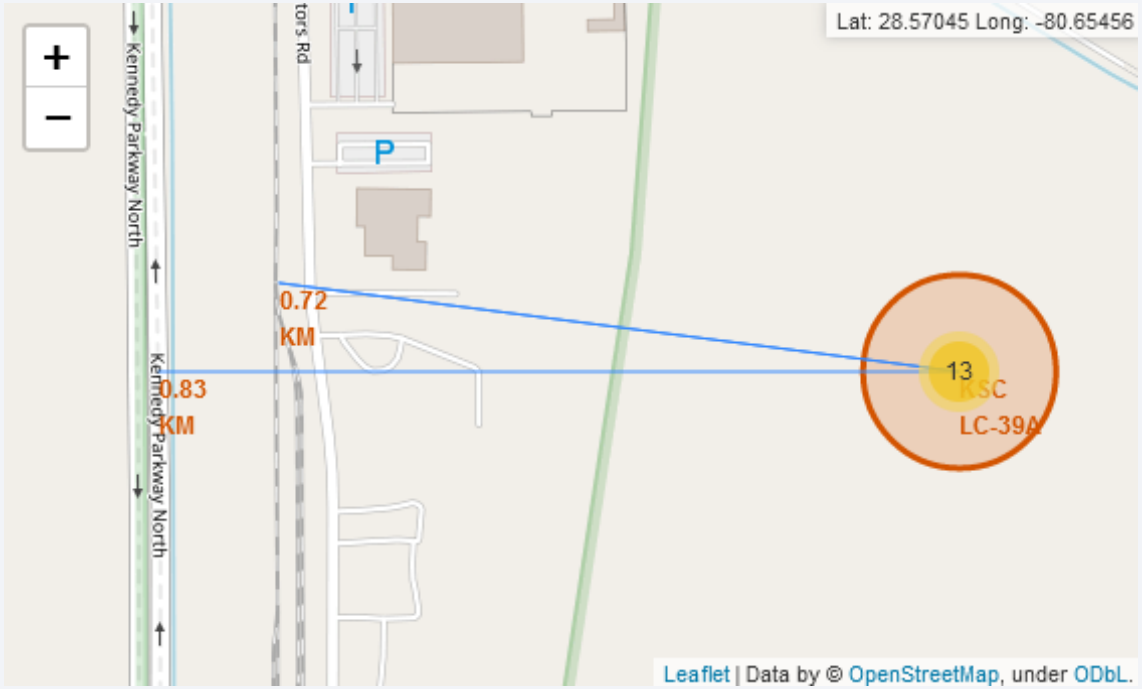
Landing outcomes out of the site (success in green, failure in red)



Proximity of site KSC LC-39A to other infrastructures

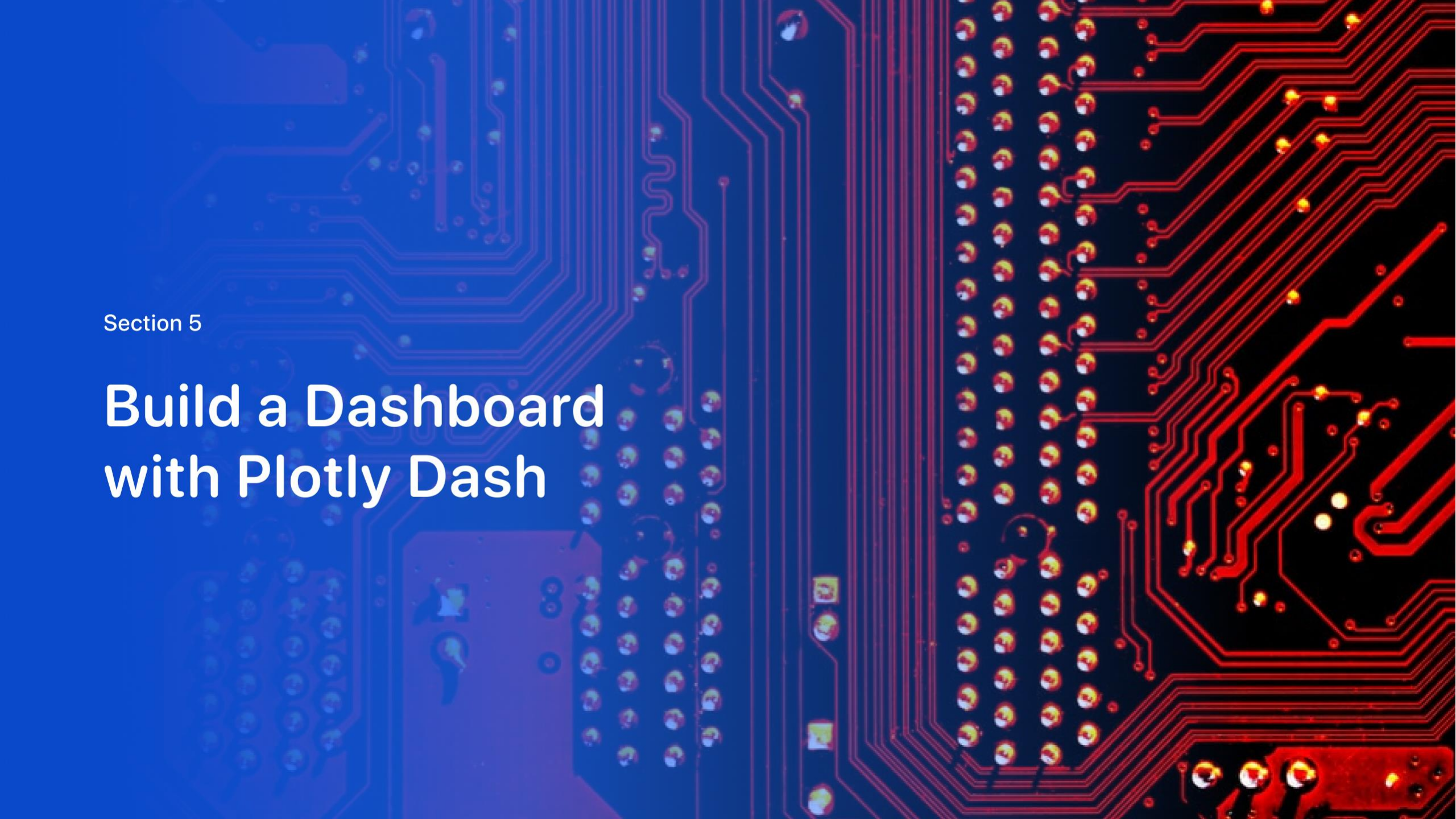
- The KSC LC-39A launch site lies in very close proximity to a railway, located less than 1 km west from the site.
- Similarly, it is also located in close proximity to a parkway, also less than 1 km to the west from the site.
- Not visible on the map, the launch site is however positioned much further away from inhabited areas, with Merritt Island located almost 19 km to the south of the launch site.

Location and proximity of closest railway and road to the site



Section 5

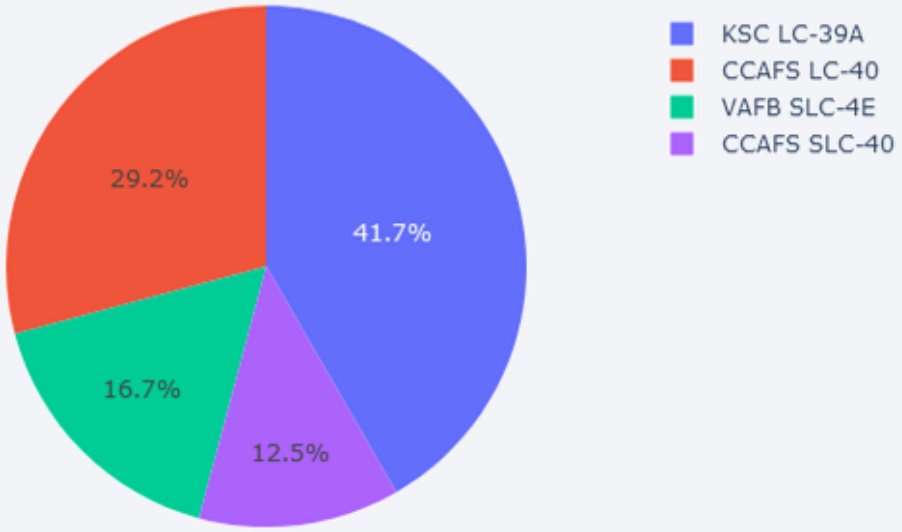
Build a Dashboard with Plotly Dash



Distribution of successful landings across launch sites

- Across all launch sites, the highest number of successful launches (i.e. launches where the first stage was successfully recovered) were from flights launched from site KSC LC-39A

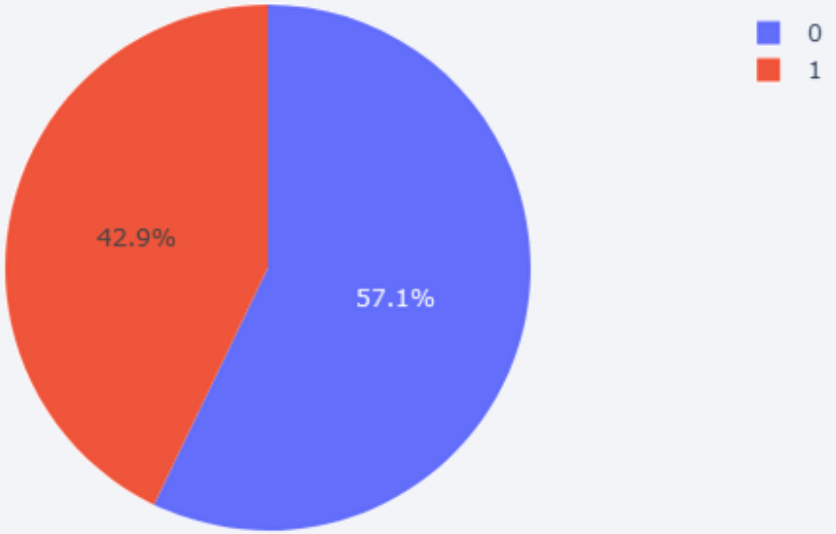
Successful launches across all sites



Mission outcome for the most successful launch site

- With a 42.9% success rate, flights launched from site CCAFS SLC-40 had the highest rate of recovery of the first stage of the launcher across all four sited used for the Falcon-9 rocket.

Mission outcome for launch site CCAFS SLC-40

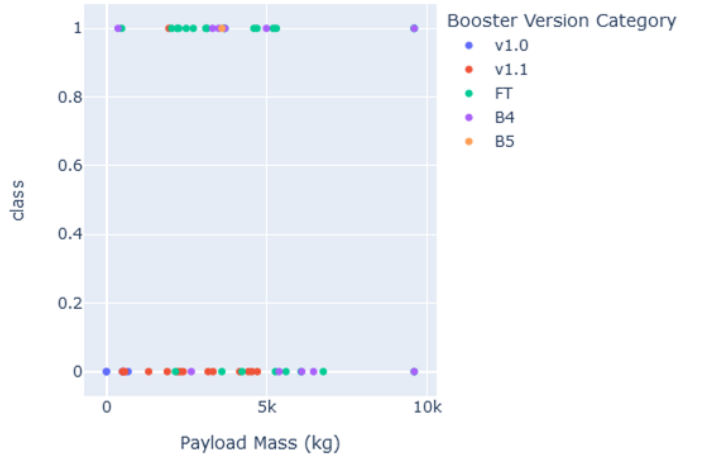


Landing success across different payload mass ranges

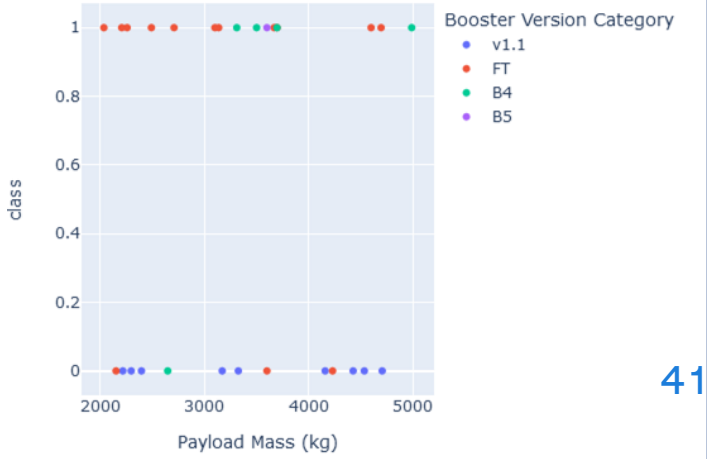
- Across all payload masses, the B5 and FT versions of the launcher seems to have the highest success rate, whereas the v1.0 and v1.1 seem to have seldom been successful.

- Looking at the payload mass, the highest success rate seems to occur between 2,000 and 5,000 kg.

Mission outcome by booster across the full range of masses



Mission outcome by booster on the range with a high success rate



Section 6

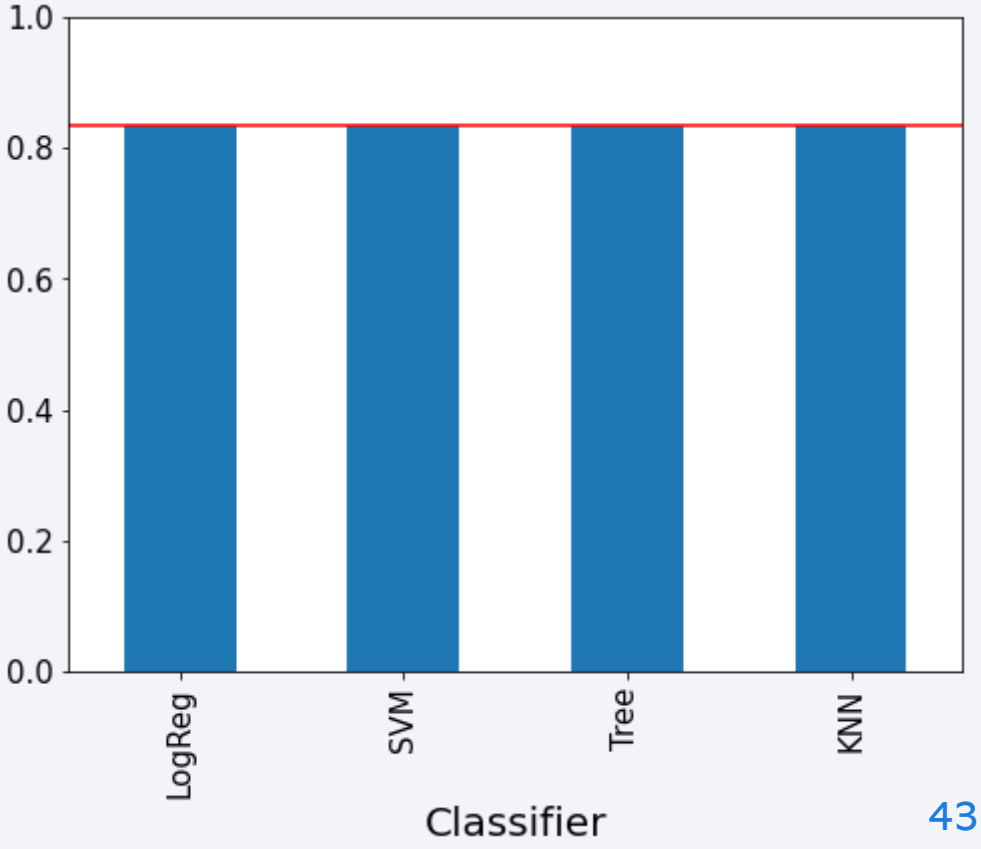
Predictive Analysis (Classification)

Classification Accuracy

Methodology

- 4 models were used to try and determine the outcome of a launch: Logistic Regression, Support Vector Machine, Decision Tree, and K-Nearest Neighbors
- After standard scaling and sub-setting the available launch data into training and test sets, finding the best hyperparameters and training the algorithms, all resulted in the same accuracy of ca. 83% on the test set.
- Considering that all the models yield the exact same predicted values, no combination of these 4 models could improve the overall accuracy above 83%.

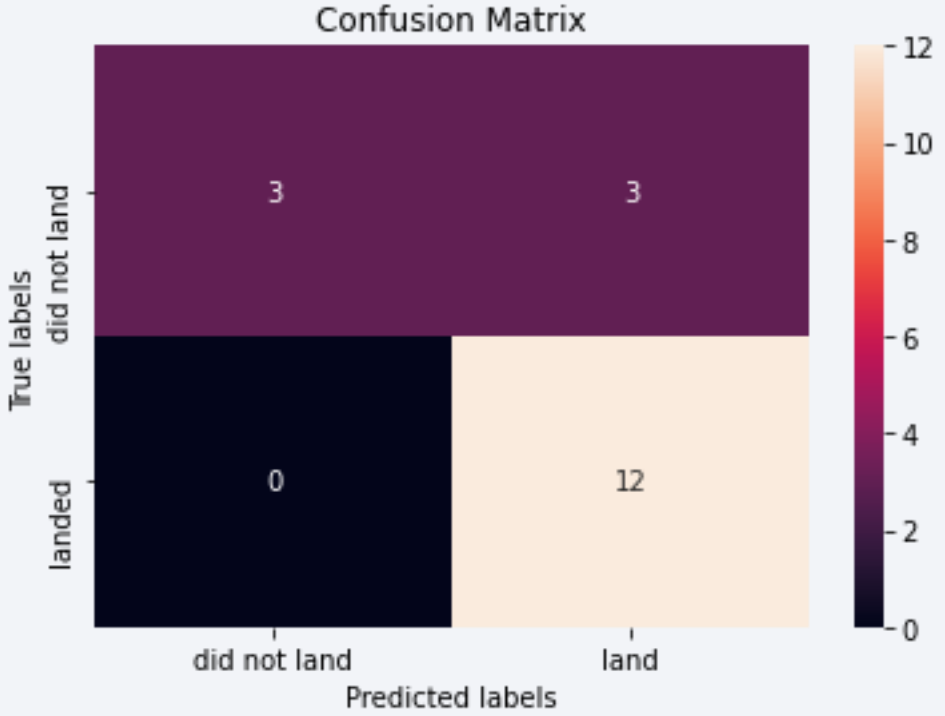
Ranked successful landing outcomes



Confusion Matrix

- As shown earlier, all models have the same accuracy. They also all show the same confusion matrix, and hence the same errors.
- Inaccuracies on the test set stemmed from false positives, where every model incorrectly predicted 15 successful launches where only 12 were reported, i.e. 3 false positives.
  - All 3 false positives were launched from launch site CCAFS SLC-40, had grid fins and legs

Confusion matrix for all models trained



Conclusions

- Based on a list of launch attributes, machine learning models can be built to predict whether the booster used will successfully land
- Using different machine learning algorithms, the accuracy of such a model can be estimated at around 83%
- All 4 models explored for this purpose have been able to accurately predict actual successful landings, with no false negatives, but seem to overstate the number of successes, leading to a number of false positives
- For competitive purposes, we can use any of the 4 models created, with a high level of confidence if a failure to land is predicted.

Appendix 1
Tuned hyperparameters for each model

Logistic Regression

- Inverse regularization strength ($C = \frac{1}{\lambda}$) = 0.01
- Penalization using ridge regression (L2)
- L-BFGS solver

Decision Tree

- Split quality measured using Gini impurity
- Maximum depth of 4 nodes
- Up to $\sqrt{n\ features}$ features considered when assessing each split
- At least 2 samples required for each leaf node
- At least 10 samples required for each internal node
- Best random split method selected

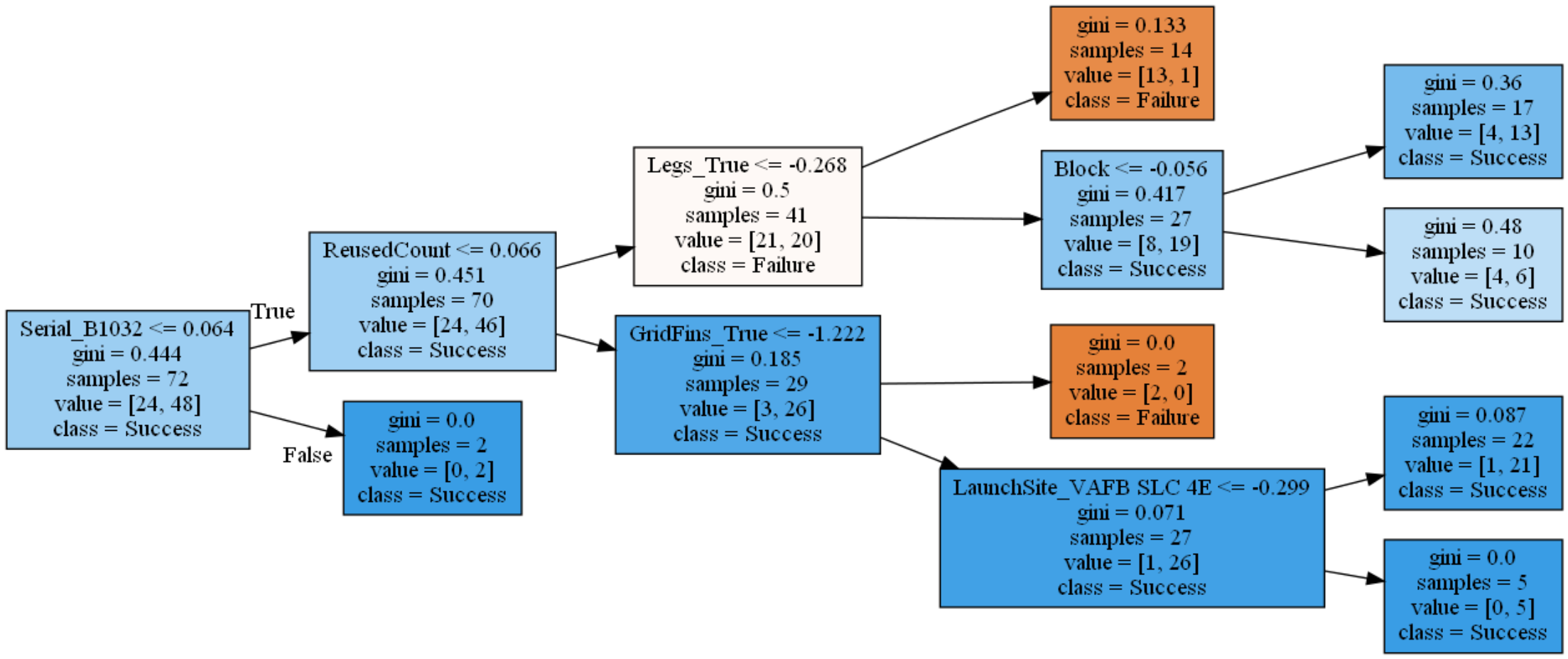
Support Vector Machine

- Regularization parameter (C) = 1.0
- Sigmoid kernel
- Kernel coefficient = 0.0316

K-Nearest Neighbors

- Automatic selection of the most appropriate algorithm to compute the nearest neighbors
- 10 neighbors in a neighborhood
- Manhattan distance used as Minkowski metric

Appendix 2
Decision tree overview



Thank you!