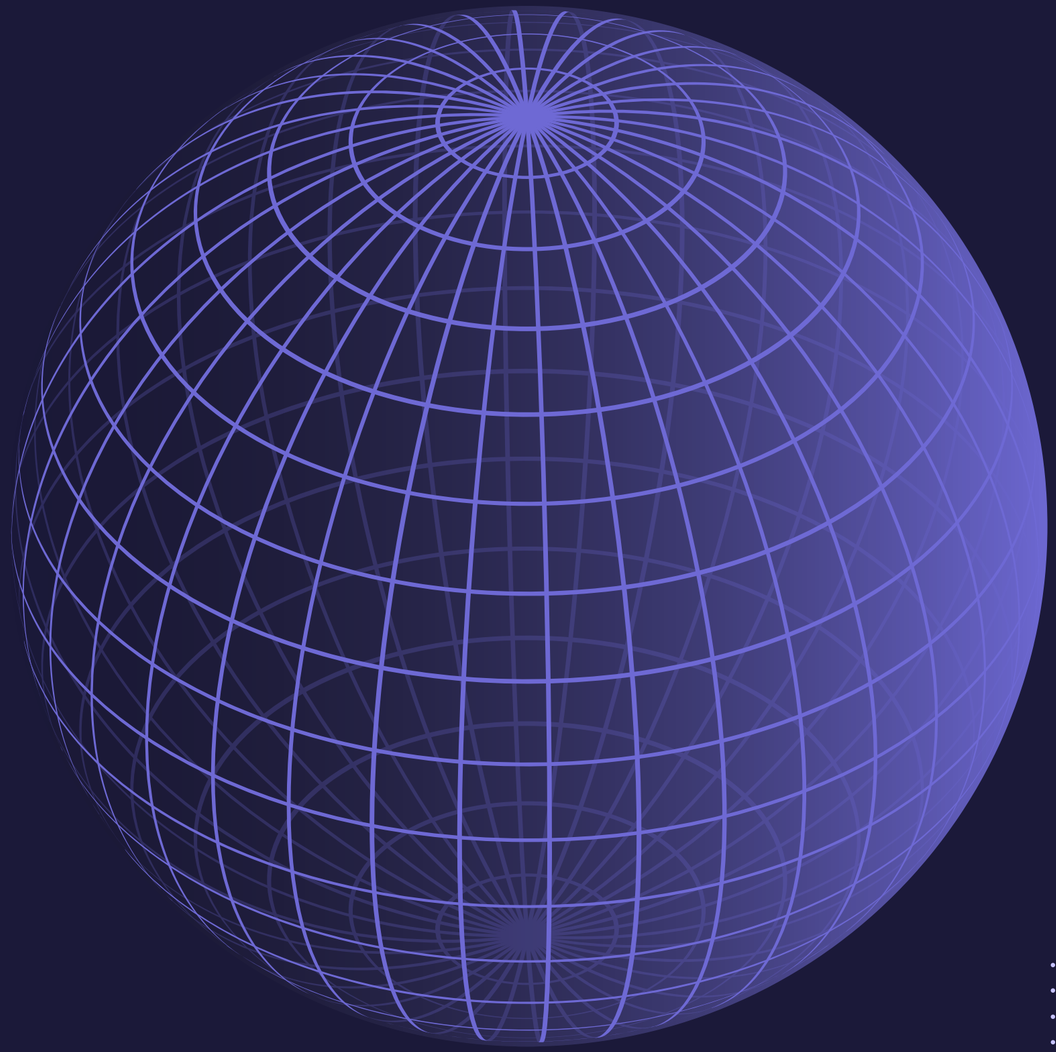
DATA SCIENCE CAPSTONE PROJECT

SAMANTHA TAN



OUTLINE

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

EXECUTIVE SUMMARY

METHODOLOGIES

This research attempts to identify the factors for a successful rocket landing. To make this determination, the following methodologies were used:

- Data collection
- Data wrangling
- Exploratory data analysis with data visualisation
- Exploratory data analysis with SQL
- Building an interactive map with Folium
- Building a dashboard with Plotly Dash
- Predictive analysis (classification)

RESULTS

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

INTRODUCTION

Background

SpaceX is the most successful company of the commercial space age, making space travel affordable. The company advertises Falcon 9 rocket launches on its website, with a cost of $62M; other providers cost upward of $165M each. This disparity is due to SpaceX's ability to reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage.

Explore

- How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?
- Does the rate of successful landings increase over the years?
- What is the best algorithm that can be used for binary classification in this case?

METHODOLOGY

DATA COLLECTION

- SpaceX Rest API
- Web scraped from Wikipedia

DATA WRANGLING

- Filtered the data
- Dealt with missing values
- One-hot-encoding to prepare the data to a binary classification

EXPLORATORY DATA ANALYSIS USING VISUALISATION AND SQL

INTERACTIVE VISUAL ANALYTICS USING FOLIUM AND PLOTLY DASH

PREDICTIVE ANALYSIS USING CLASSIFICATION MODELS

- Building, tuning, and evaluation of classification models to ensure the best results

DATA COLLECTION

Data collection involved a combination of API requests from SpaceX REST API and web scraping data from a table in SpaceX's Wikipedia entry.

Both of these data collection methods were used to get complete information about the launches for a more detailed analysis.

DATA COLUMNS OBTAINED USING SPACEX REST API:

- FlightNumber
- Date
- BoosterVersion
- Payload Mass
- Orbit
- LaunchSite
- Outcome
- Serial
- Flights
- GridFins
- Reused
- Legs
- LandingPad
- Block
- ReusedCount
- Latitude
- Longitude

DATA COLUMNS OBTAINED USING WIKIPEDIA WEB SCRAPING:

- Flight No.
- Launch site
- Payload
- PayloadMass
- Orbit
- Customer
- Launch outcome
- Version Booster
- Booster landing
- Date
- Time

DATA COLLECTION - SPACEX API

1. Requested rocket launch data from SpaceX API
2. Decoded the response content using *.json()* and turned it into a dataframe using *.json_normalize()*
3. Requested needed information about the launches from SpaceX API by applying custom functions
4. Constructed the data obtained into a dictionary
5. Created a dataframe from the dictionary
6. Filtered the dataframe to only include Falcon 9 launches
7. Replaced missing values of Payload Mass with calculated *.mean()*
8. Exported the data to CSV

DATA COLLECTION - WEB SCRAPING

1. Requested Falcon 9 launch data from Wikipedia
2. Created a BeautifulSoup object from the HTML response
3. Extracted all column names from the HTML table header
4. Collected the data by parsing HTML tables
5. Constructed the data obtained into a dictionary
6. Created a dataframe from the dictionary
7. Exported the data to CSV

DATA WRANGLING

In the data set, there were several different cases where the booster did not land successfully. Some landings were attempted but failed due to an accident. The outcomes of these missions were interpreted as follows:

True Ocean: The mission successfully landed to a specific region of the ocean

False Ocean: The mission unsuccessfully landed to a specific region of the ocean

True RTLS: The mission successfully landed to a ground pad

False RTLS: The mission unsuccessfully landed to a ground pad

True ASDS: The mission successfully landed on a drone ship

False ASDS: The mission unsuccessfully landed on a drone ship

These outcomes were converted to training labels with ‘1’ representing a successful landing, and ‘0’ representing an unsuccessful landing.

PERFORM EXPLORATORY DATA ANALYSIS AND DETERMINED TRAINING LABELS

- Calculated the number of launches from each site
- Calculated the number and occurrence of each orbit
- Calculated the number and occurrence of mission outcomes per orbit type
- Created a landing outcome label from the Outcome column
- Exported the data to CSV

EXPLORATORY DATA ANALYSIS WITH DATA VISUALISATION

The following charts were plotted:

1. Flight Number vs. Payload Mass
2. Flight Number vs. Launch Site
3. Payload Mass (kg) vs. Launch Site
4. Orbit Type vs. Success Rate
5. Flight Number vs. Orbit Type
6. Payload Mass (kg) vs. Orbit Type
7. Success Rate Yearly Trend

Scatter plots show the relationship between variables. If a relationship exists, they could be used in machine learning models.

Bar chart show comparisons among discrete categories. The goal is to show the relationship between the specific categories.

Line charts show trends in data over time.

PERFORMED SQL QUERIES

Displayed:

- Names of the unique launch sites
- 5 records where launch sites begin with the string ‘CCA’
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1

Listed:

- Date of the first successful landing outcome on a ground pad
- Names of the boosters that had success in landing on a drone ship, and had a payload mass that was greater than 4,000 but less than 6,000
- Total number of successful and failed mission outcomes
- Names of the booster versions that carried the maximum payload mass
- Failed landing outcomes on a drone ship, their booster versions, and launch site names for the months in 2015
- Count of landing outcomes between the date 2010-06-04 and 2017-03-20 in descending order

EXPLORATORY DATA ANALYSIS WITH SQL

INTERACTIVE VISUAL ANALYTICS WITH FOLIUM

Markers of all Launch Sites

- Added s blue circle at NASA Johnson Space Center with a popup label and text label showing its name using its latitude and longitude coordinates
- Added red circles at all launch sites with popup labels and text labels showing their names using their latitude and longitude coordinates

Coloured Markers of the Launch Outcomes for each Launch Site

- Added coloured markers of successful (green) and unsuccessful (red) launches using *MarkerCluster()* to identify which launch sites have relatively high success rates

Distances Between a Launch Site to its Proximities

- Added coloured lines to show distances between the launch sites and its proximities to the nearest railways, highways, coastlines, and city

INTERACTIVE VISUAL ANALYTICS WITH PLOTLY DASH

IBM DATA SCIENCE CAPSTONE | SAMANTHA TAN

LAUNCH SITES DROPDOWN LIST

- Added a dropdown list to enable launch site selection

PIE CHART SHOWING SUCCESSFUL LAUNCHES

- Added a pie chart to show the total successful launches count for all sites, and the Success vs. Failed counts for the site if a specific launch site was selected

SLIDER OF PAYLOAD MASS RANGE

- Added a slider to select payload range

SCATTER CHART OF PAYLOAD MASS VS. SUCCESS RATE FOR THE DIFFERENT BOOSTER VERSIONS

- Added a scatter chart to show the correlation between payload and launch success

PREDICTIVE ANALYSIS (CLASSIFICATION)

1. Created a NumPy array from the column ‘Class’ in data
2. Standardised the data with *StandardScaler()*, then fitted and transformed it
3. Split the data into training and testing sets with *train_test_split()*
4. Created a GridSearchCV object with cv = 10 to find the best parameters
5. Applied *GridSearchCV()* on LogReg, SVM, Decision Tree, and KNN models
6. Calculated the accuracy on the test data using *.score()* for all models
7. Examined the confusion matrices for all models
8. Found the best performing model by examining the Jaccard and F1 Scores

RESULTS

EXPLORATORY DATA ANALYSIS

- Launch success has improved over time
- KSC LC-39A has the highest success rate among landing sites
- Orbits ES-L1, GEO, HEO, and SSO have a 100% success rate

VISUAL ANALYTICS

- Most launch sites are near the equator, and all are close to the coast
- Launch sites are far away enough from anything a failed launch can damage (city, highway, railway), while still close enough to bring people and materials to support launch activities

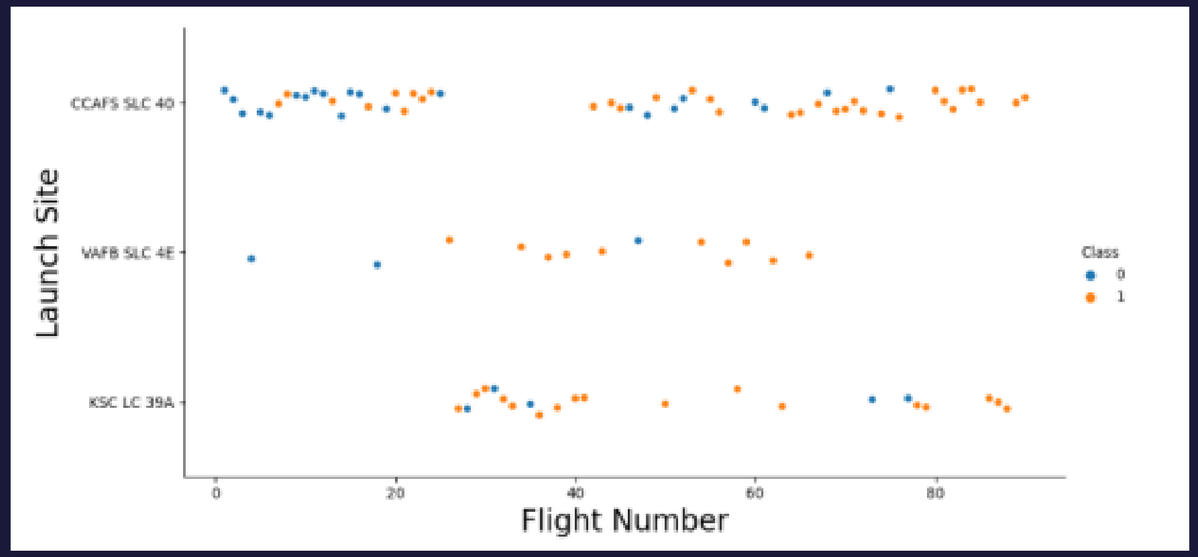
PREDICTIVE ANALYSIS

Decision Tree model is the best predictive model for the data set

EXPLORATORY DATA ANALYSIS

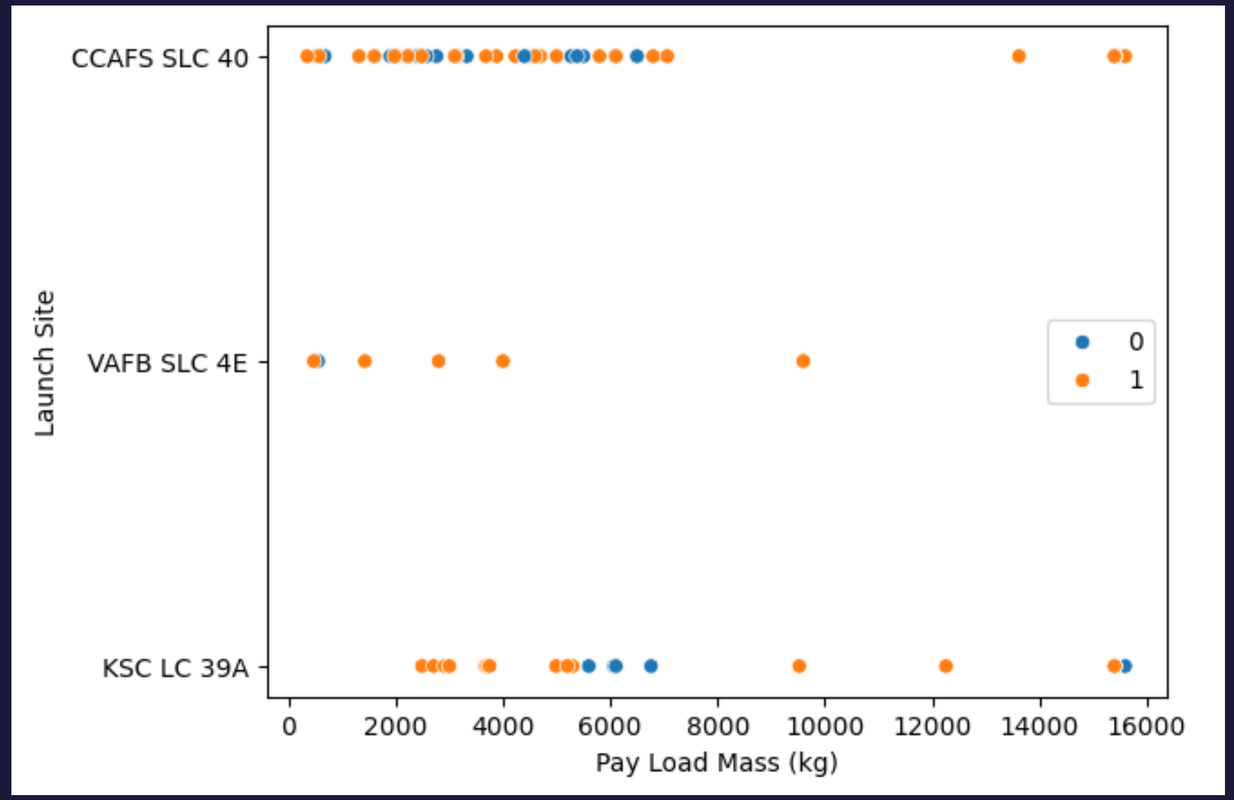
IBM DATA SCIENCE CAPSTONE | SAMANTHA TAN

FLIGHT NUMBER VS. LAUNCH SITE



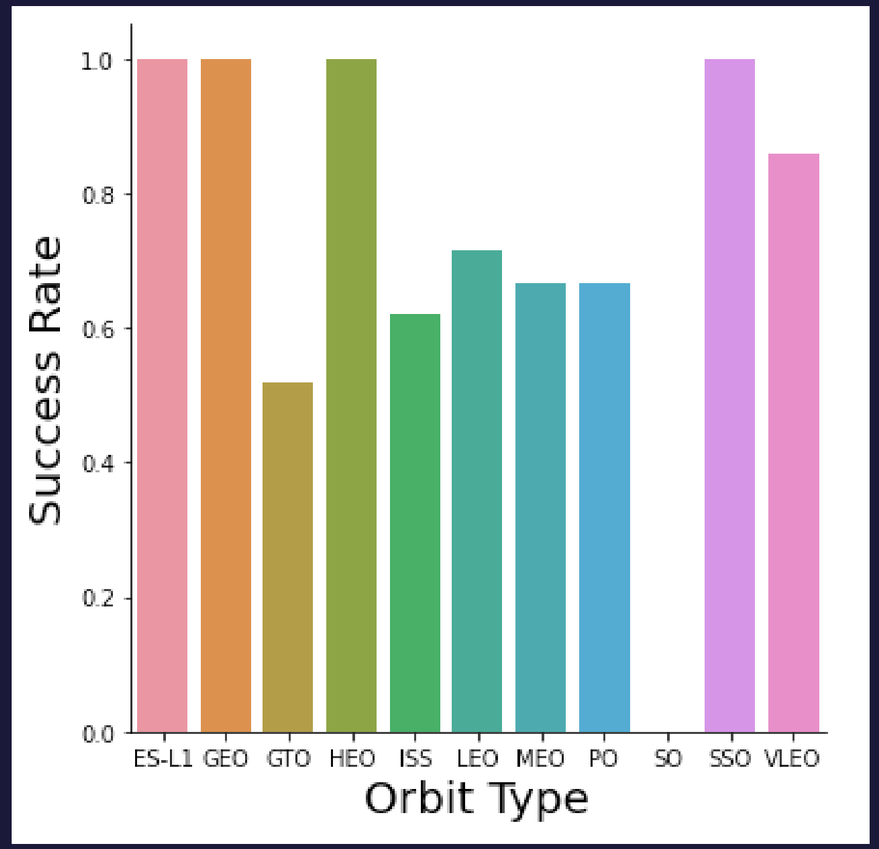
- Earlier flights had a lower success rate, while later flights had a higher success rate
- About half of all launches were from the CCAFS SLC-40 launch site
- VAFB SLC-4E and KSC LC-39A have higher success rates
- New launches are likely to have a higher success rate

PAYLOAD MASS (KG) VS. LAUNCH SITE



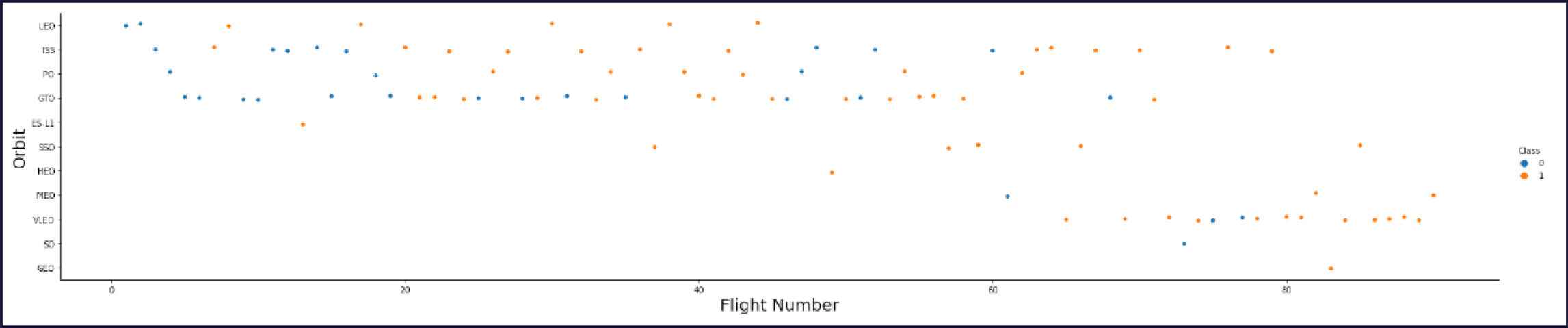
- Generally, the higher the payload mass, the higher the success rate
- Most of the launches with payload mass over 7,000kg were successful
- KSC LC-39A has a 100% success rate for payload mass under 5,500kg
- VAFB SLC-4E has not launched anything greater than 10,000kg

SUCCESS RATE VS. ORBIT TYPE



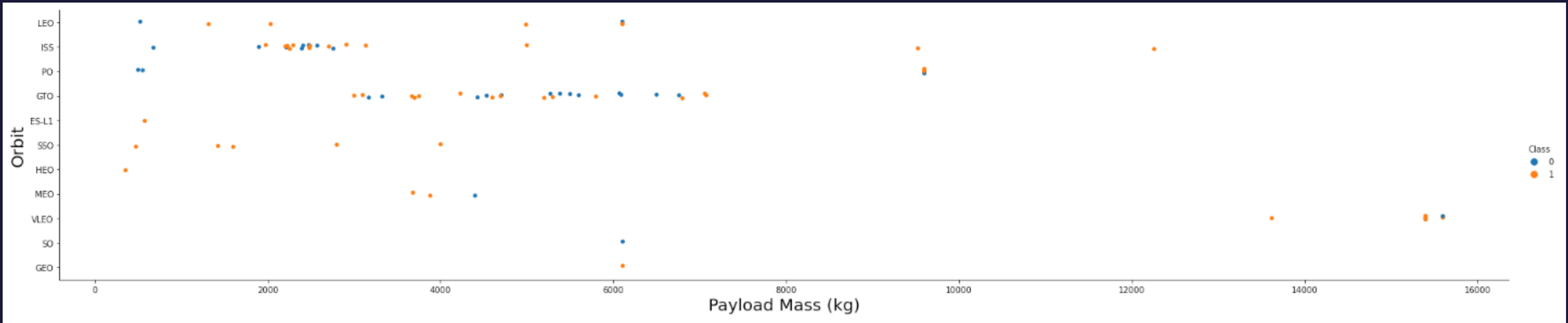
- Orbits with 100% success rate:
  - ES-L1
  - GEO
  - HEO
  - SSO
- Orbits with success rates between 50% and 85%:
  - GTO
  - ISS
  - LEO
  - MEO
  - PO
- Orbits with 0% success rate:
  - SO

FLIGHT NUMBER VS. ORBIT TYPE



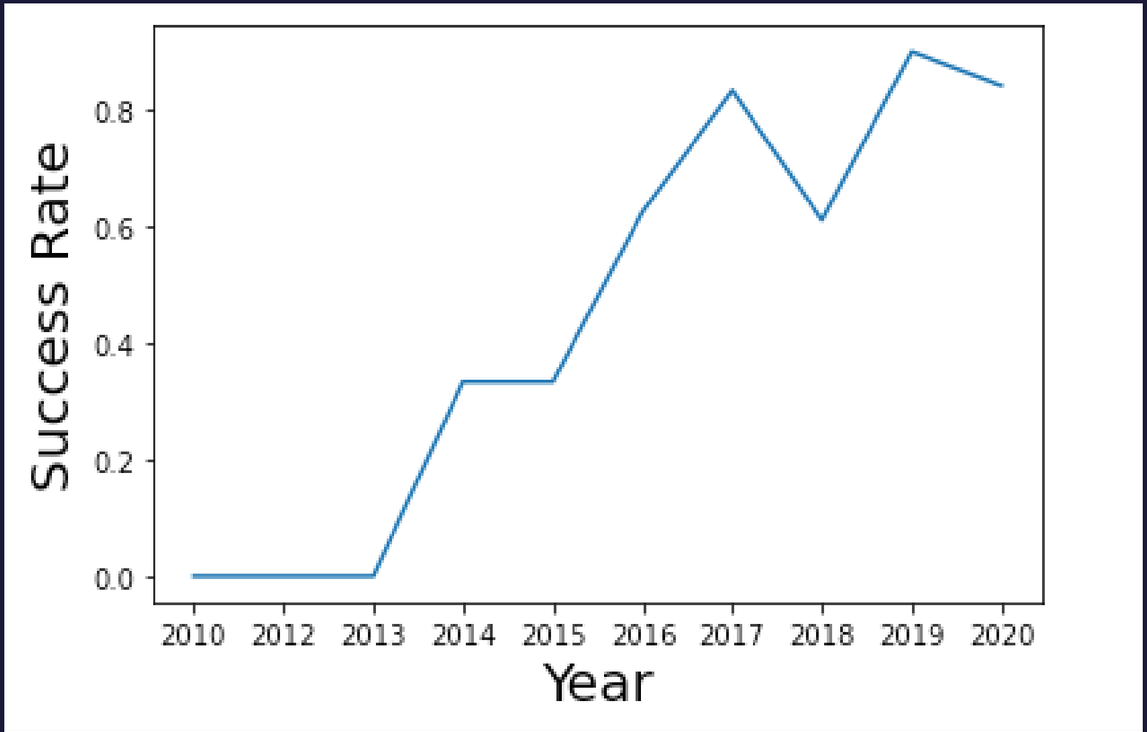
- The success rate increases with the number of flights for each orbit
- In the LEO orbit, the success appears to be related to the number of flights
- In the GTO orbit, there seems to be no relationship with flight number

PAYLOAD MASS (KG) VS. ORBIT TYPE



- Heavy payloads have a positive influence on the LEO, ISS, and PO orbits;
- Heavy payloads have no influence on the GTO orbit

LAUNCH SUCCESS YEARLY TREND



- The success rate improved from 2013 - 2017 and 2018 - 2019
- The success rate decreased from 2017 - 2018 and 2019 - 2020
- However, the success rate has improved since 2013 overall

LAUNCH SITE INFORMATION

Launch Sites

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

5 Records with Launch Site Starting 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

PAYLOAD MASS

TOTAL PAYLOAD MASS

45, 596 kg carried by boosters launched by NASA (CRS)

TotalPayloadMass
45596

AVERAGE PAYLOAD MASS

2,534.7 kg carried by booster version F9 v1.1

AveragePayloadMass
2534.6666666666665

LANDING & MISSION INFORMATION

1ST SUCCESSFUL LANDING IN GROUND PAD

22/12/2015

FirstSuccessfulLandingDate
2015-12-22

BOOSTER DRONE SHIP LANDING

- Booster mass greater than 4,000 but less than 6,000
- JSCAT-14, JSCAT-16, SES-10, SES-11 / EchoStar 105

JCSAT-14
JCSAT-16
SES-10
SES-11 / EchoStar 105

TOTAL NUMBER OF SUCCESSFUL AND FAILED MISSION OUTCOMES

- 1 Failure (in flight)
- 99 Success
- 1 Success (payload status unclear)

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

BOOSTERS

CARRYING MAX PAYLOAD

- 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

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

FAILED LANDINGS ON DRONE SHIP

In 2015

Showing month, date, booster version, launch site, and landing outcome

month	Date	Booster_Version	Launch_Site	Landing _Outcome
01	10-01-2015	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	14-04-2015	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

SUCCESSFUL LANDINGS

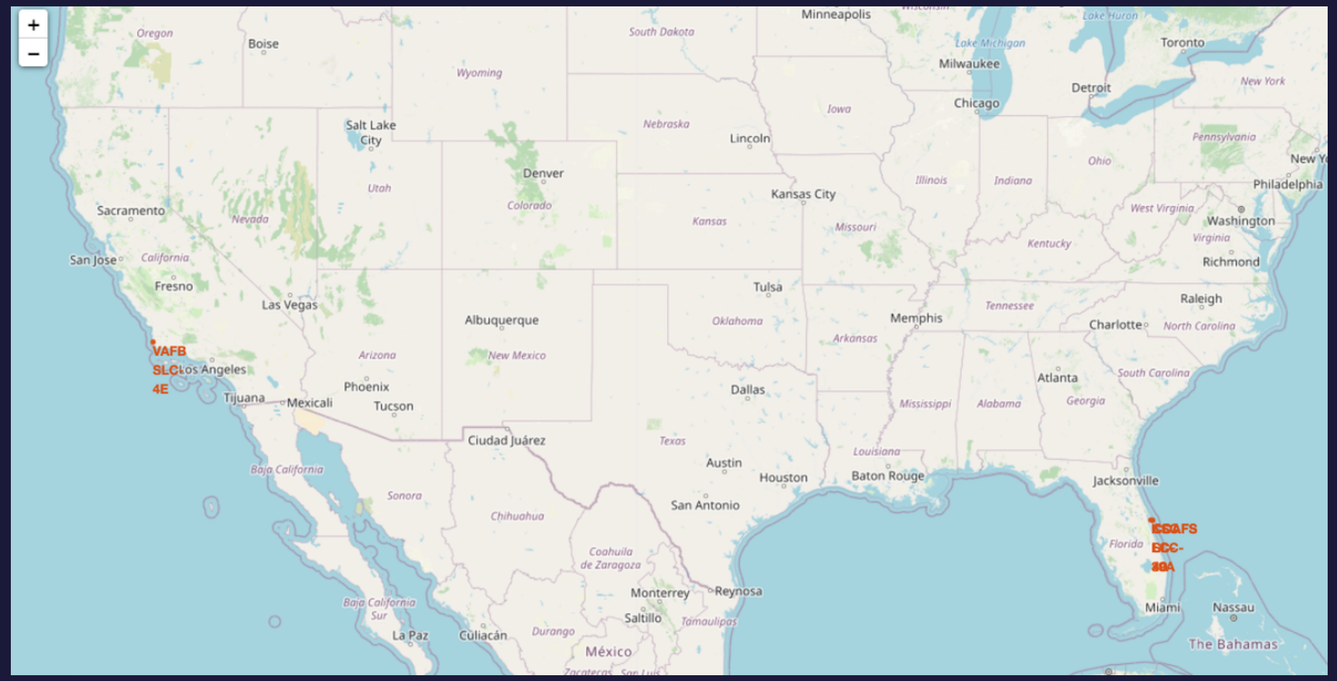
COUNT OF LANDING OUTCOMES BETWEEN 2010-06-04 AND 2017-03-20 IN DESCENDING ORDER

Landing_Outcome	Count_Landings
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

IBM DATA SCIENCE CAPSTONE | SAMANTHA TAN

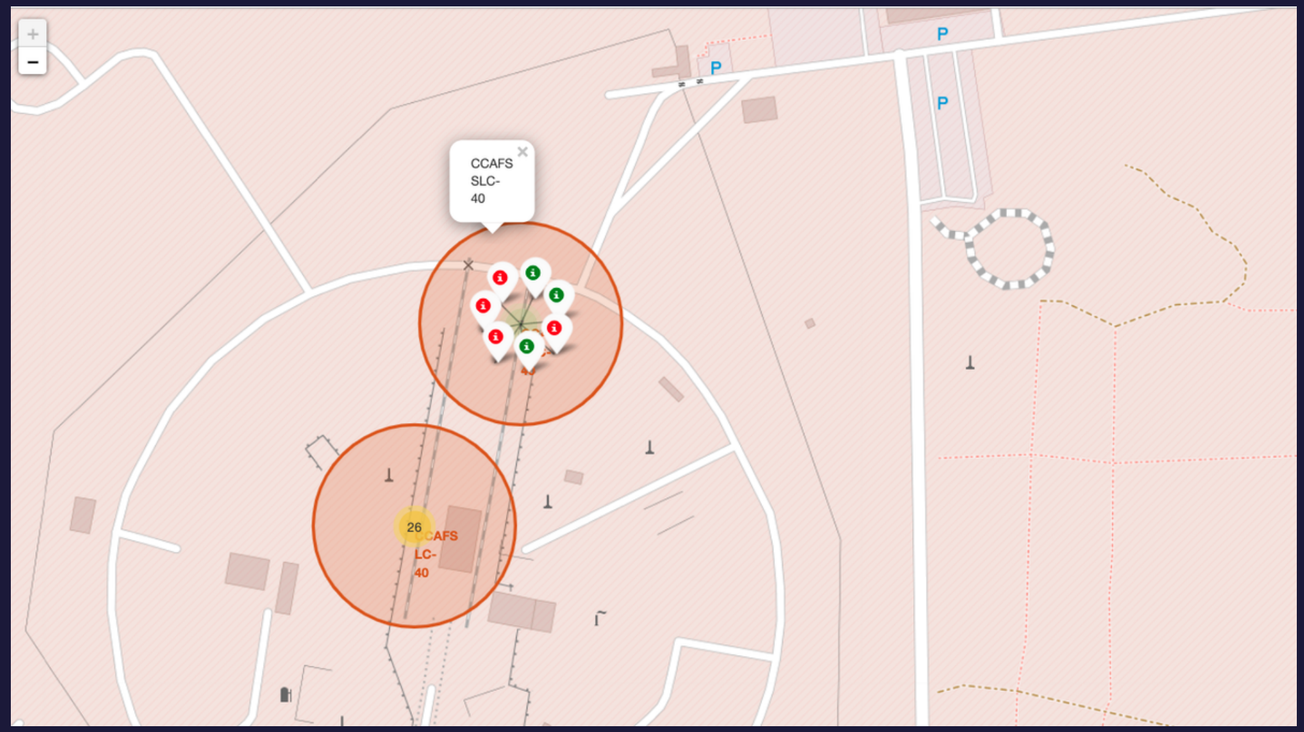
VISUAL ANALYTICS

LAUNCH SITES



- The launch sites are near the equator
- The closer they are to the equator, the easier it is to launch to equatorial orbit, and the more help the launches get from Earth’s rotation for a prograde orbit
- This helps save the cost of putting in extra fuel and boosters

LAUNCH OUTCOMES



- Green markers: successful launches
- Red markers for unsuccessful launches
- Launch site CCAFS SLC-40 has a 42.9% success rate

DISTANCE TO PROXIMITIES

CCAFS SLC-40:

- 0.86 km from nearest coastline
- 21.96 km from nearest railway
- 23.23 km from nearest city
- 26.88 km from nearest highway

DISTANCE TO PROXIMITIES

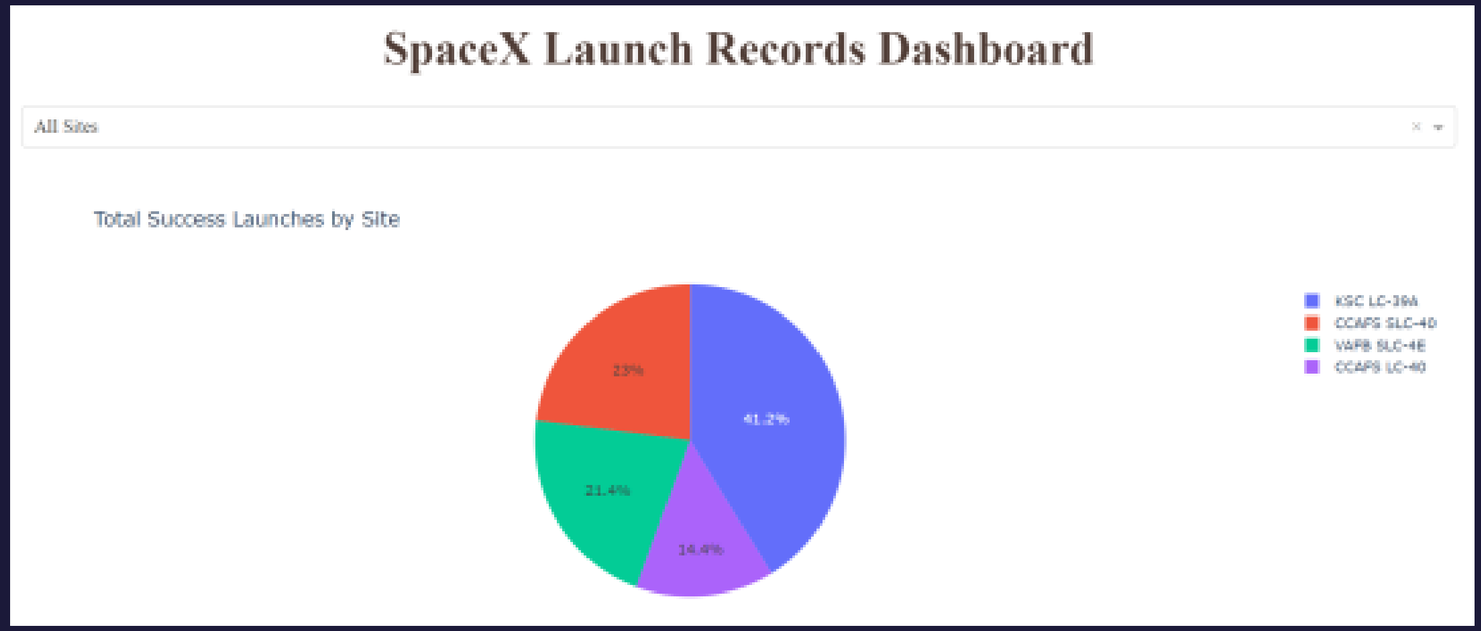
CCAFS SLC-40:

- **Coasts:** help ensure that spent stages dropped along the launch path or failed launches don't fall on people or property

- **Safety & Security:** there needs to be an exclusion zone around the launch site to keep unauthorised people away and safe

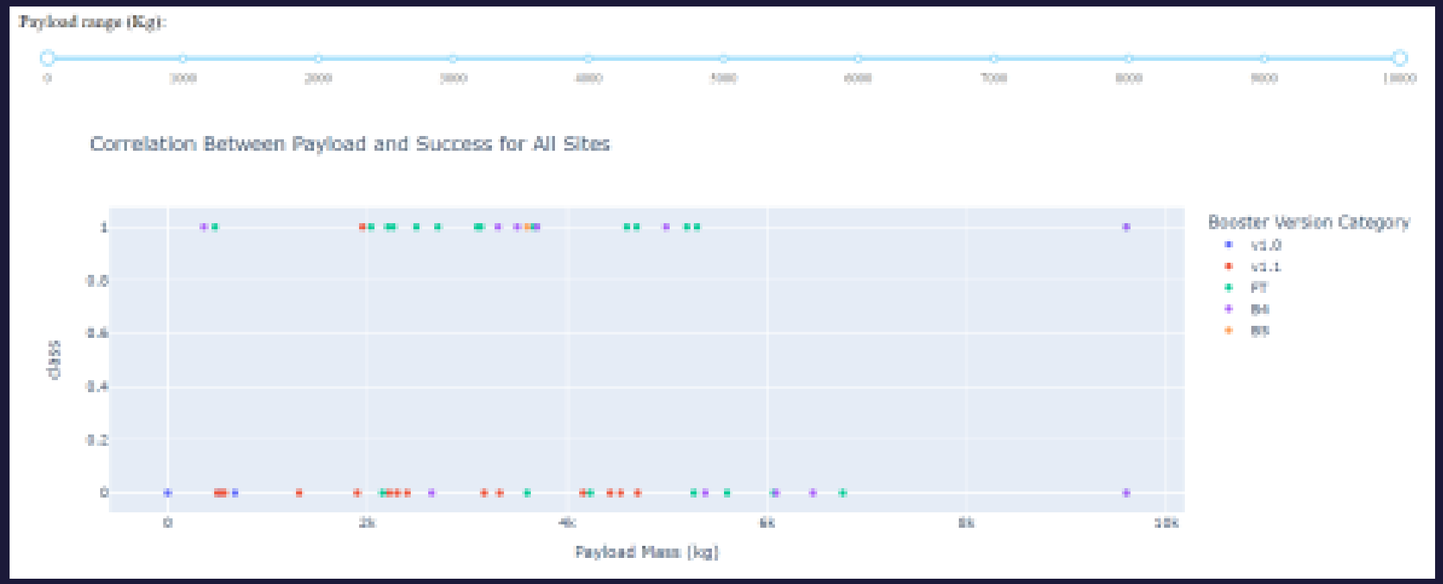
- Transportation, Infrastructure, and Cities: need to be away from anything a failed launch can damage, but close enough to roads/rails/docks to be able to bring people and materials to or from it in support of launch activities

LAUNCH SUCCESS BY SITE



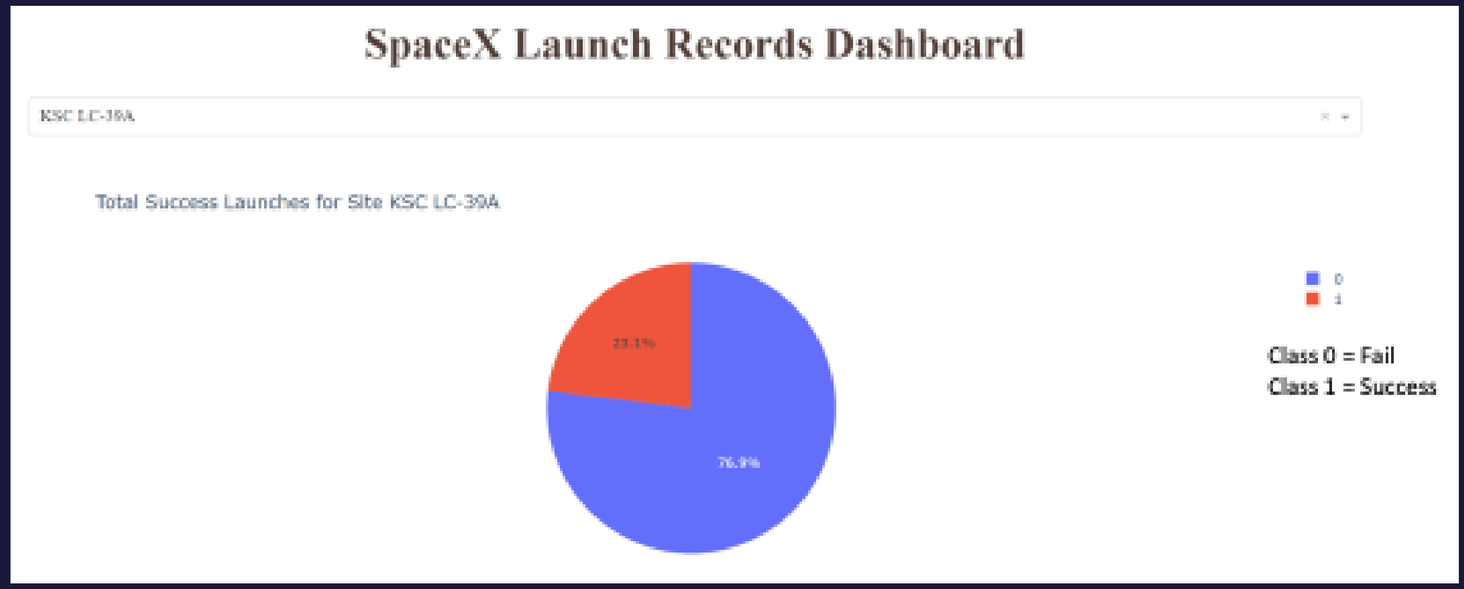
- Success as Percent of Total
- KSC LC-39A has the most successful launches amongst launch sites (41.2%)

PAYLOAD MASS AND SUCCESS



- By Booster Version
- Payloads between 2,000 kg and 5,000 kg have the highest success rate
- ‘1’ indicates successful outcomes, and ‘0’ indicates unsuccessful outcomes

LAUNCH SUCCESS - KSC LC-39A



- Success as Percent of Total
- KSC LC-39A has the most successful launches amongst launch sites (76.9%)
- 10 successful launches, and 3 failed launches

IBM DATA SCIENCE CAPSTONE | SAMANTHA TAN

PREDICTIVE ANALYSIS

CLASSIFICATION

Accuracy

- All the models performed at about the same level and had the same scores and accuracy
- This is likely due to the small dataset
- The Decision Tree model slightly outperformed the rest when looking at *.best_score_*
- *.best_score_* is the average of all cv folds for a single combination of parameters

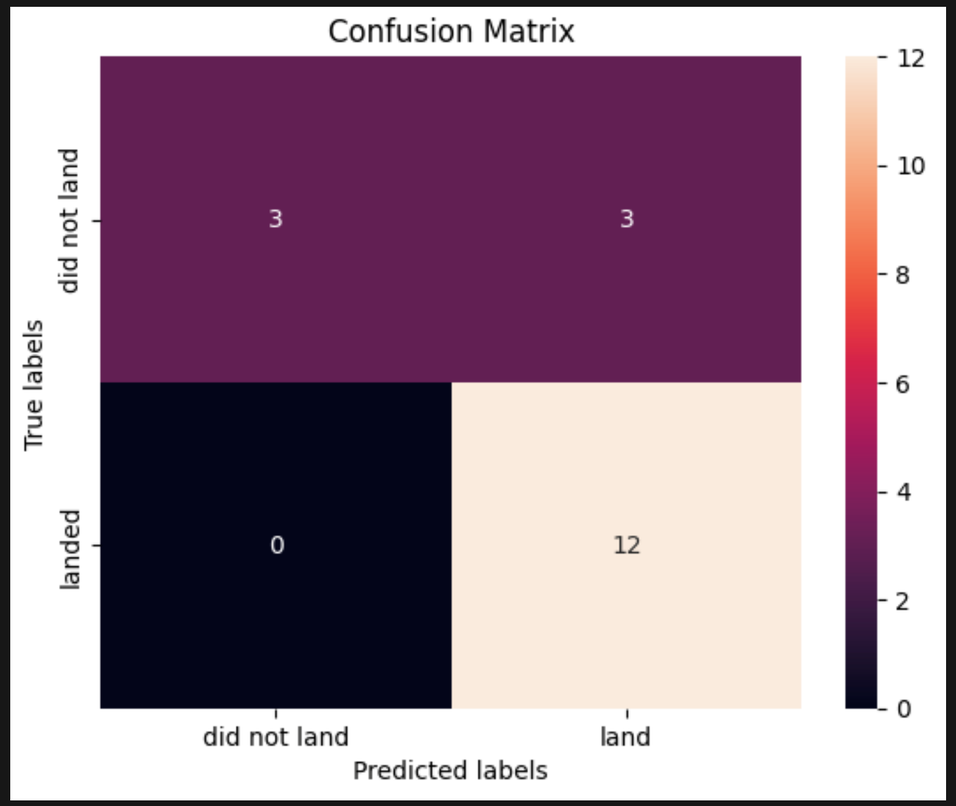
CONFUSION MATRICES

Performance Summary

- A confusion matrix summarises the performance of a classification algorithm
- All the confusion matrices were identical
- The presence of false positives (Type 1 Error) is not good

- Confusion Matrix Outputs:
  - 12 True Positive
  - 3 True Negative
  - 3 False Positive
  - 0 False Negative

- Precision = 0.80
- Recall = 1
- F1 Score = 0.89
- Accuracy = 0.833



CONCLUSION

RESEARCH

Model Performance	The models performed similarly on the test set with the decision tree model slightly outperforming
Equator	Most of the launch sites are near the equator for an additional natural boost - due to the rotational speed of Earth - which helps save the cost of putting in extra fuel and boosters
Coast	All the launch sites are close to the coast
Launch Success	Increases over time
KSC LC-39A	Has the highest success rate among launch sites; Has a 100% success rate for launches less than 5,000 kg
Orbits	ES-L1, GEO, HEO, and SSO have a 100% success rate
Payload Mass	Across all launch sites, the higher the payload mass (kg), the higher the success rate

CONCLUSION

THINGS TO CONSIDEER

Dataset	A larger dataset will help build on the predictive analytics results to help understand if the findings can be generalisable to a larger data set
Feature Analysis / PCA	Additional feature analysis or principal component analysis should be conducted to see if it can help improve accuracy
XGBoost	Is a powerful model which was not utilised in this study. It would be interesting to see if it outperforms the other classification models

THANK YOU