IBM
DATA SCIENCE CAPSTONE

MOHAMED SHAHEEN
29-08-2024

OUTLINE

Executive summary

Introduction

Methodology

Results

Conclusion

Appendix

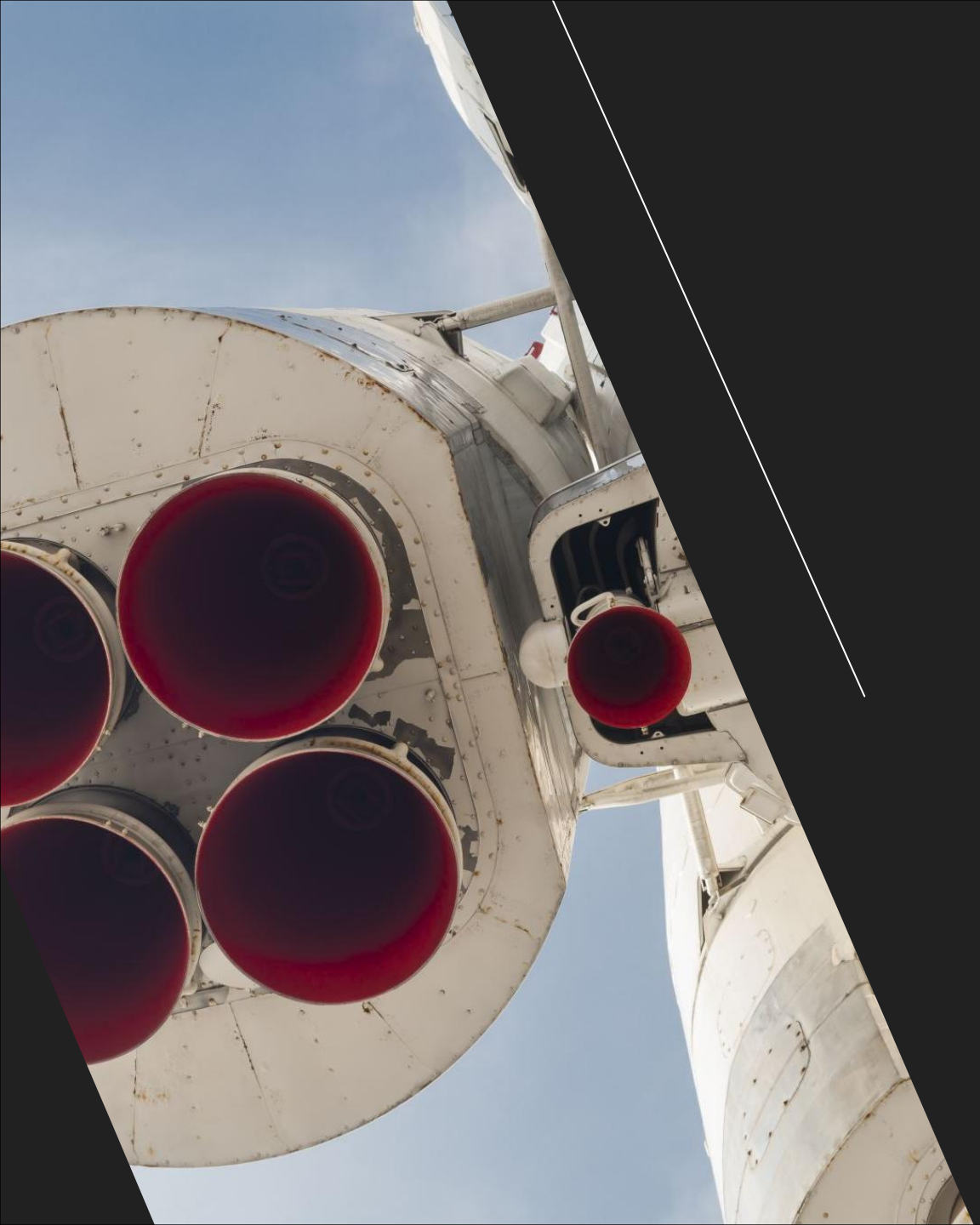
EXECUTIVE SUMMARY

Summary of all results

- Data collection
- Data wrangling
- Exploratory Data Analysis with Data Visualization
- Exploratory Data Analysis with SQL
- Building an interactive map with Folium
- Predictive analysis (Classification)

Summary of methodologies

- Exploratory Data Analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results



INTRODUCTION

SpaceX is a pioneering force in the commercial space industry, known for significantly reducing the costs of space travel. The Falcon 9 rocket, advertised at $62 million per launch, offers a more affordable option compared to competitors, who charge upwards of $165 million. This cost reduction is primarily due to the reuse of the rocket's first stage. The project's goal is to predict the likelihood of this first-stage reuse based on various factors, utilizing machine learning models and publicly available data.

KEY QUESTIONS:

- How do variables like payload mass, launch site, and orbit type influence the success of the first-stage landing?
- Has the success rate of landings improved over time?
- Which classification algorithm is the most effective for this prediction task?

METHEDOLOGY

- **Data Collection:**

Employed the SpaceX REST API and web scraping from Wikipedia to gather comprehensive data on launches.

- **Data Wrangling:**

Filtered and cleaned the data, handled missing values, and used One-Hot Encoding to prepare for analysis.

- **Exploratory Data Analysis (EDA):**

Visualized data relationships and trends using charts and SQL queries to derive insights.

- **Interactive Visualization:**

Used Folium for creating a geographical map of launch sites and Plotly Dash to build an interactive dashboard.

- **Predictive Analysis:**

Implemented and fine-tuned various classification models (Logistic Regression, SVM, Decision Tree, KNN) to predict landing success, evaluating them with accuracy metrics.

Methodology

DATA COLLECTION

The data collection process utilized both API requests from the SpaceX REST API and web scraping from SpaceX's Wikipedia page to gather comprehensive information on rocket launches. Using both methods was essential to ensure complete and detailed data for analysis.

Data Columns obtained from the SpaceX REST API include:

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

Data Columns obtained through Wikipedia Web Scraping include:

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

DATA COLLECTION- Space X API

- Requesting rocket launch data from Space X API
- Decoding the response content using .json() and turning it into a data frame using .json_normalize()
- Requesting essential info about the launches from Space x API
- Constructing the data we obtained into a dictionary
- Creating a dataframe from a dictionary
- Filtering the data frame to only include datas from Falcon9 launches
- Replacing the missing values of Payload Mass column with calculated mean of this coloumn
- Exporting data to csv

DATA COLLECTION-WEB SCARPING

1. Requesting Falcon 9 launch data from Wikipedia
2. Creating a BeautifulSoup object from the HTML response
3. Extracting all column names from the HTML table header
4. Collecting the data by parsing HTML tables
5. Constructing the obtained data into a dictionary
6. Creating a dataframe from the dictionary
7. Exporting the data to CSV

DATA WRANGLING

In the dataset, there are various instances where the booster did not land successfully. Sometimes, a landing was attempted but failed due to an accident. For example, "True Ocean" indicates that the mission successfully landed in a specific region of the ocean, while "False Ocean" indicates that the mission failed to land in the intended ocean region. Similarly, "True RTLS" means the mission successfully landed on a ground pad, whereas "False RTLS" means the landing on the ground pad was unsuccessful. "True ASDS" signifies a successful landing on a drone ship, while "False ASDS" indicates an unsuccessful landing on a drone ship.

These outcomes are primarily converted into training labels: "1" denotes a successful booster landing, and "0" denotes an unsuccessful landing.

- Perform exploratory Data Analysis and determine Training Labels
- Calculate the number of launches on each site
- Calculate the number and occurrence of each orbit
- Calculate the number and occurrence of mission outcome per orbit type
- Create a landing outcome label from Outcome column
- Exporting the data to CSV

EDA WITH DATA VISUALIZATION

Charts plotted

Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type, Payload Mass vs Orbit Type and Success Rate Yearly Trend

- Scatter plots show the relationship between variables. If a relationship exists, they could be used in machine learning model.
- Bar charts show comparisons among discrete categories. The goal is to show the relationship between the specific categories being compared and a measured value.
- Line charts show trends in data over time (time series).

EDA WITH SQL

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string ‘CCA
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.
- Listing the date when the first successful landing outcome in ground pad was achieved
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster versions which have carried the maximum payload mass
- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
- Ranking 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

MAP WITH FOLIUM

- Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.
- Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts.
- Added coloured Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.
- Added coloured Lines to show distances between the Launch Site KSC LC-39A (as an example) and its proximities like Railway, Highway, Coastline and Closest City.

BUILD DASHBOARD WITH PLOTLY DASH

- Added a dropdown list to enable Launch Site selection.
- 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 specif ic Launch Site was selected.
- Added a slider to select Payload range.
- Added a scatter chart to show the correlation between Payload and Launch Success.

PREDICTIVE ANALYSIS

- Creating a NumPy array from the column “Class” in data
- Standardizing the data with standard scalar, then fitting and transforming it
- Splitting the data into training and testing with train test split function
- Creating gridsearchcv object with cv = 10 to find the best parameters
- Applying GridSearchCV on LogReg, SVM, Decision Tree, and KNN models
- Calculating the accuracy on the test data using the method .score() for all models
- Examining the confusion matrix for all models
- Finding the method performs best by examining the Jaccard_score and F1_score metrics

RESULTS

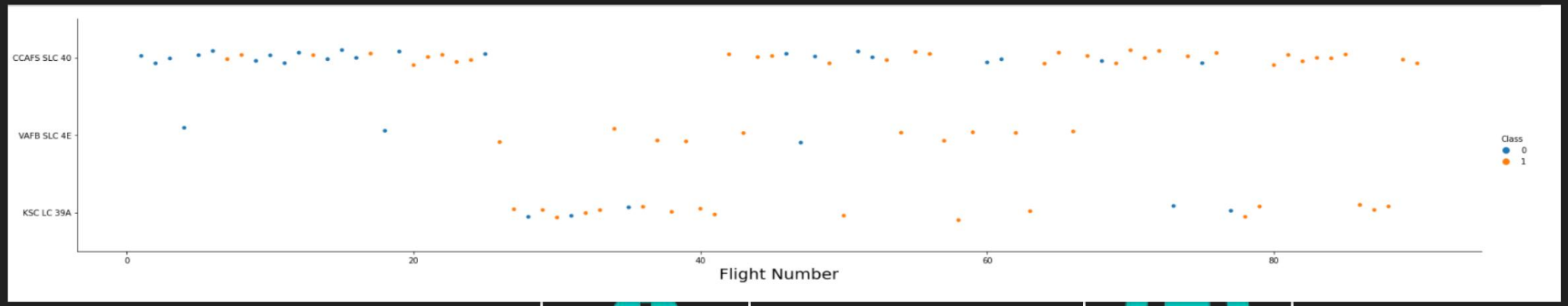
Exploratory data analysis results

Interactive analytical demo

Predictive analysis results

EDA WITH VISUALIZATION

FLIGHT NUMBER VS LAUNCH SITE



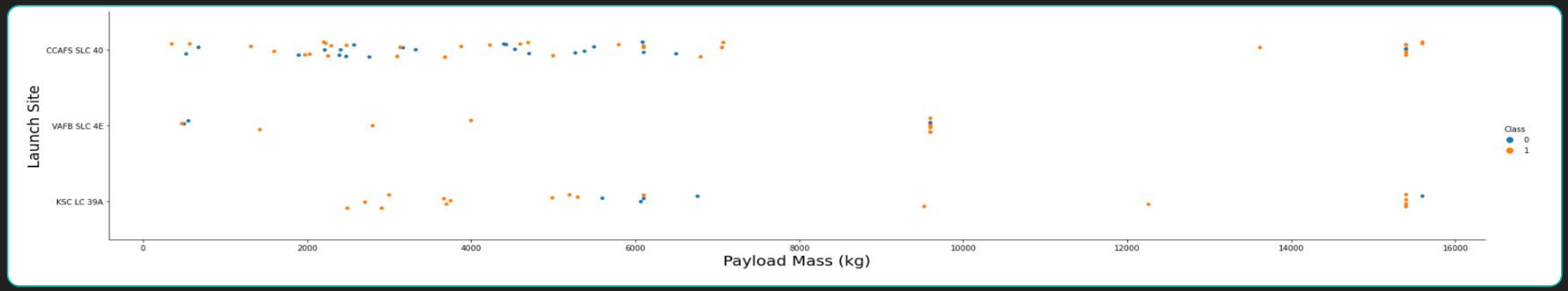
The earliest flights all failed while the latest flights all succeeded

VAFB SLC 4E and KSC LC 39A have higher success rates.

The CCAFS SLC 40 launch site has about a half of all launches

It can be assumed that each new launch has a higher rate

PAYLOAD VS LAUNCH SITE



For every launch site the higher the payload mass, the higher the success rate.

Most of the launches with payload mass over 7000 kg were successful

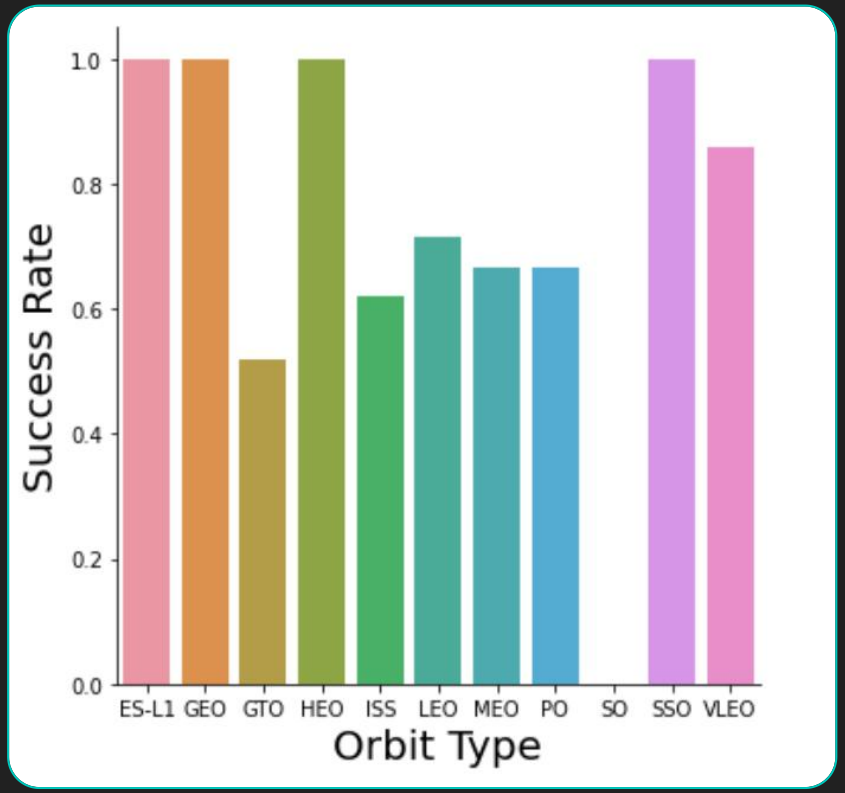
KSC LC 39A has a 100% success rate for payload mass under 5500 kg too

SUCCESS RATE VS ORBIT TYPE

Orbits with 100% success rate:
- ES-L1, GEO, HEO, SSO

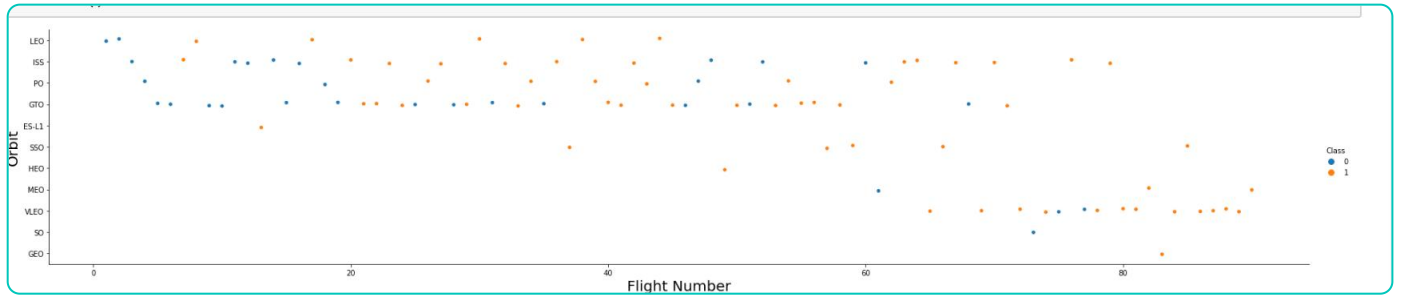
Orbits with 0% success rate:
- SO

Orbits with success rate between 50% and 85%:
- GTO, ISS, LEO, MEO, PO

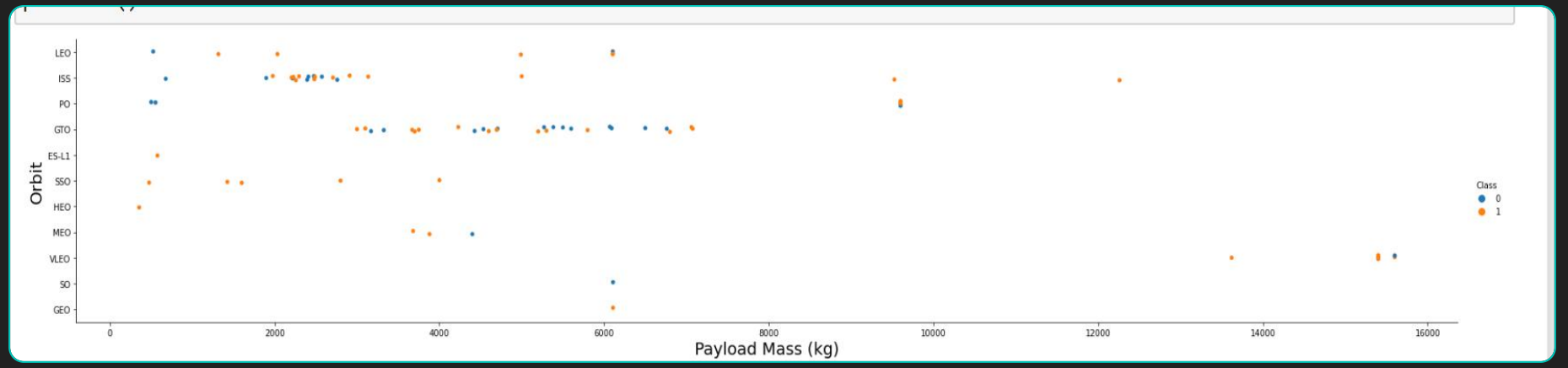


FLIGHT NUMBER VS ORBIT

- In the LEO orbit the Success appears related to the number of f lights; on the other hand, there seems to be no relationship between f light number when in GTO orbit.



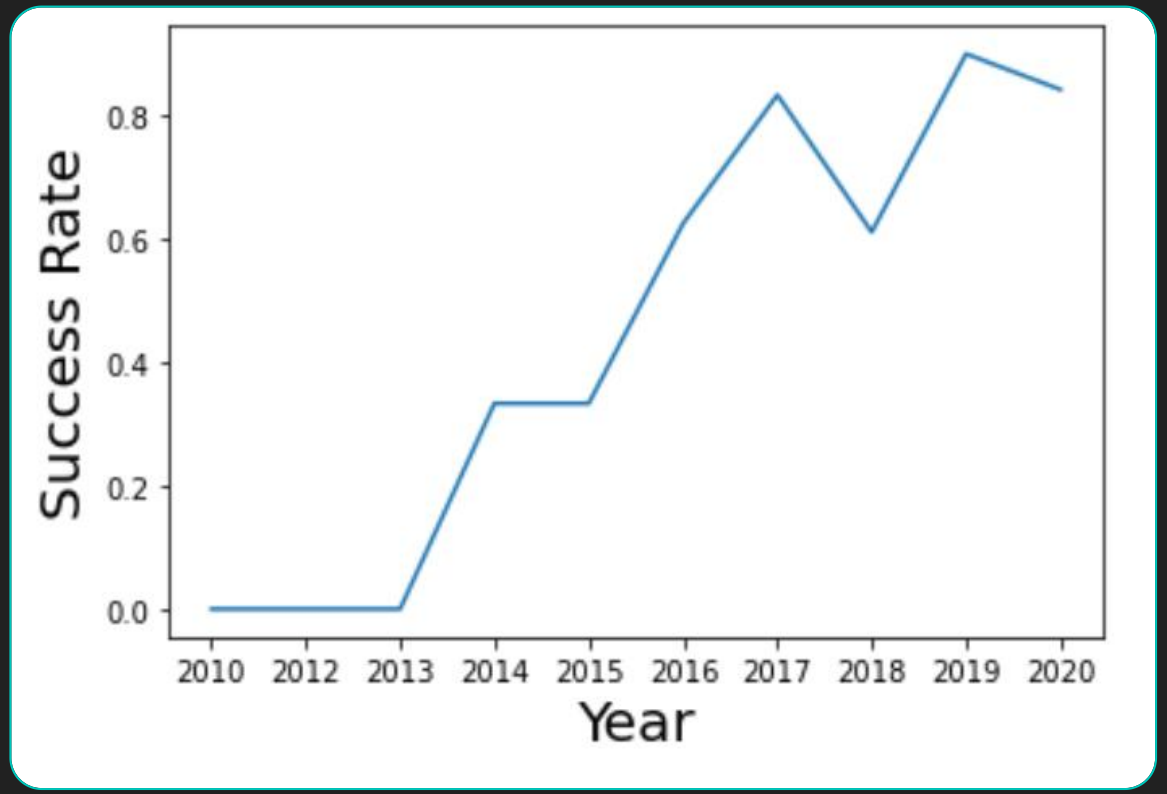
PAYLOAD VS ORBIT



- Heavy payloads have a negative inf luence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.

LAUNCH SUCCESS YEARLY TREND

- The success rate since 2013 kept increasing till 2020



EDA WITH SQL

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

ALL LAUNCH SITE NAME

Displaying the names of the unique launch sites in the space mission.

LAUNCH SITE NAME BEGIN WITH 'CAA'

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 is : 45596
- Average payload mass by F9 v1.1 : 2534
- First successful ground landing date : 2015-12-22

PAYLOAD MASS

SUCCESSFUL DRONE SHIP LANDING WITH PAYLOAD BETWEEN 4000 AND 6000

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

TOTAL NUMBER OF SUCCESSFUL FAILURE MISSION OUTCOMES

mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

BOOSTERS CARRIED MAXIUMUM PAYLOADS

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 RECORD

- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015.

MONTH	DATE	booster_version	launch_site	landing__outcome
January	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

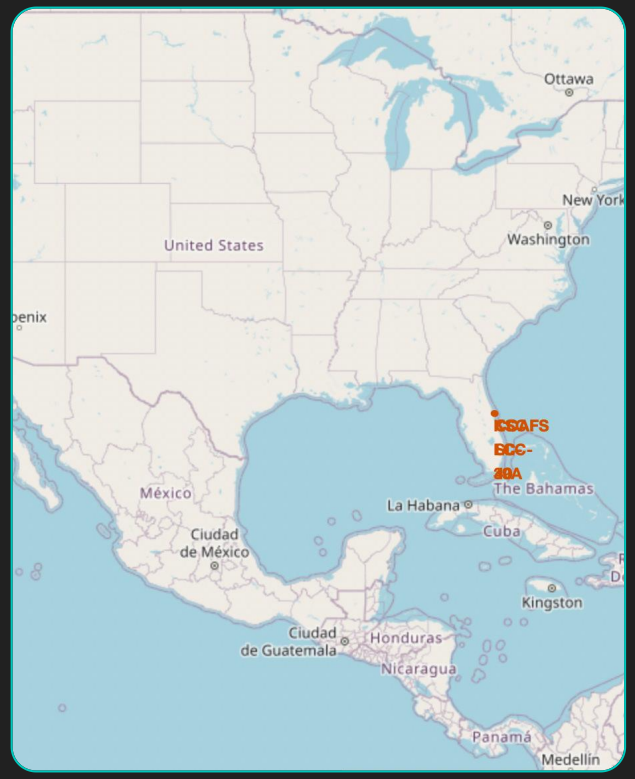
RANK SUCCESS COUNT BETWEEN 2010-06-04 & 2017-03-20

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

INTERACTIVE MAPS WITH FOLIUM

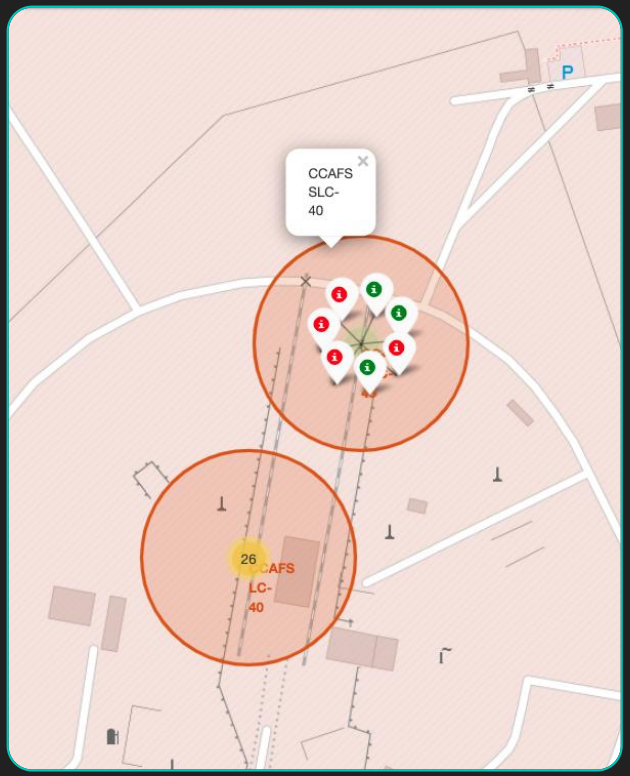
ALL LAUNCH SITE'S LOCATION MARKERS

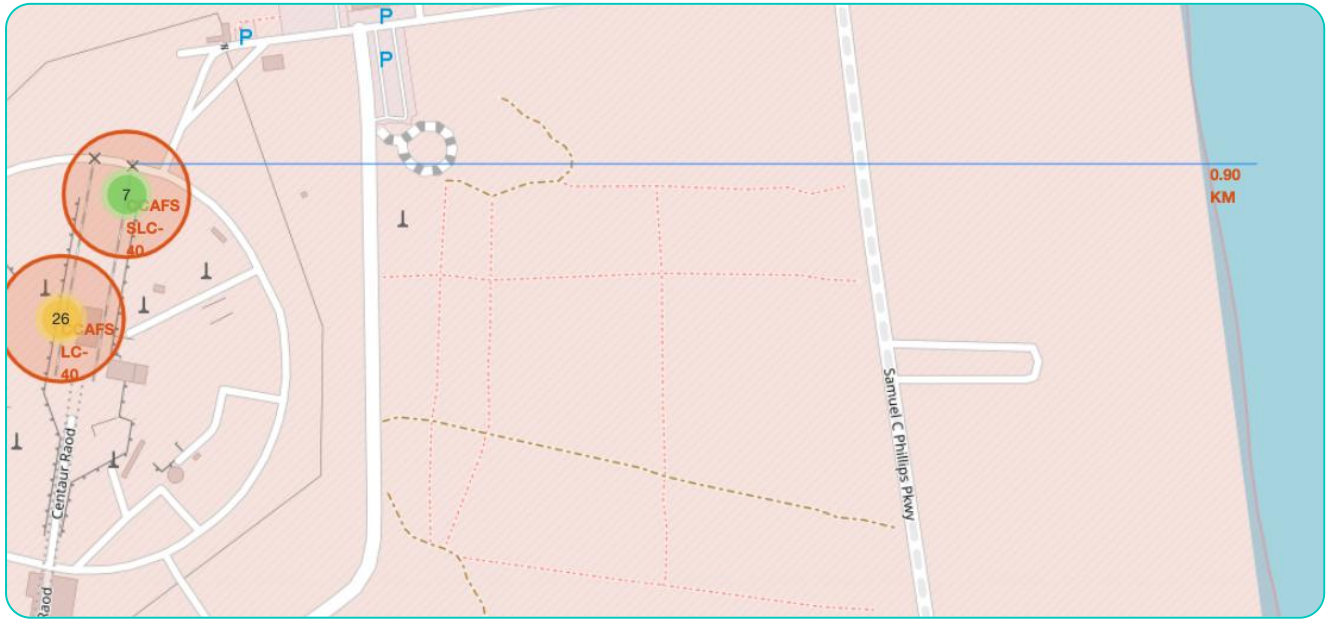


- Most launch sites are located near the equator because the land moves faster at the equator than at any other location on Earth's surface. At the equator, objects are already moving at a speed of 1,670 km/hour due to Earth's rotation. When a spacecraft is launched from the equator, it ascends into space while retaining this rotational speed, thanks to inertia. This initial speed provides a significant advantage, helping the spacecraft maintain the velocity needed to stay in orbit.
- All launch sites are located very close to the coast. Launching rockets over the ocean minimizes the risk of debris falling or exploding near populated areas, ensuring greater safety for people.

COLOUR LABEL LAUNCH RECORD ON MAP

- From the colour-labeled markers we should be able to easily identify which launch sites have relatively high success rates.
- - Green Marker = Successful Launch
- Red Marker = Failed Launch
- Launch Site KSC LC-39A has a very high Success Rate.





DISTANCE FROM LAUNCH SITE AND PROXIMITIES

PREDICTIVE ANALYSIS

CLASSIFICATION ACCURACY

Based on the scores of the Test Set, we can not confirm which method performs best.

Same Test Set scores may be due to the small test sample size (18 samples). Therefore, we tested all methods based on the whole Dataset.

The scores of the whole Dataset confirm that the best model is the Decision Tree Model. This model
has not only higher scores, but also the highest accuracy.

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

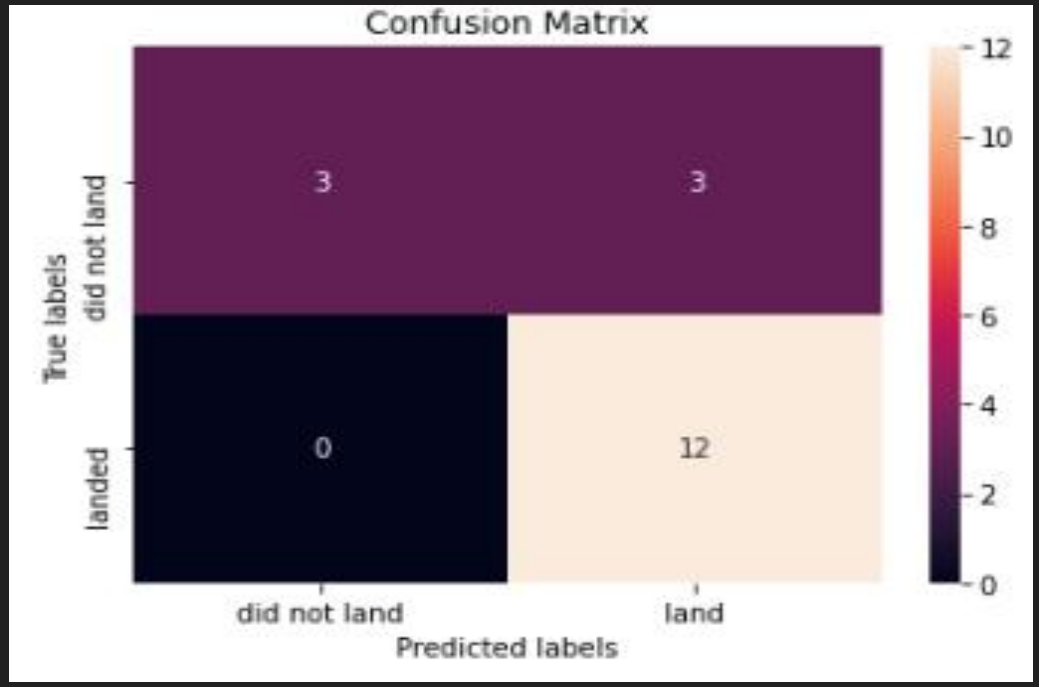
The scores from Test sets

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

The scores from the whole Dataset

CONFUSSION MATRIX

Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the major problem is false positives



CONCLUSION

- Decision Tree Model is the best algorithm for this dataset
- Launches with a low payload mass show better results than launches with a larger payload mass.
- Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast
- The success rate of launches increases over the years
- KSC LC-39A has the highest success rate of the launches from all the sites.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate

SPECIAL THANKS TO

Coursera and IBM