SpaceX Falcon 9 Rocket Recovery Prediction

This Coursera Data Science Project focuses on leveraging machine learning techniques to predict the success of Stage 1 recovery for the SpaceX Falcon 9 rocket. The ability to accurately forecast the recovery of this critical rocket component has significant implications for SpaceX's operations, enabling the company to optimize their launch procedures and reduce costs.

The project began with the collection of data from public SpaceX API and Wikipedia sources, providing a comprehensive dataset to work with. The team then conducted an extensive exploratory data analysis (EDA) process, utilizing SQL queries, visualization tools, Folium maps, and interactive dashboards to uncover insights and patterns within the data.

A key step in the project was the creation of a classification column named 'class' to label successful and unsuccessful rocket landings. This allowed the team to train and evaluate various machine learning models, identifying the most relevant features for predicting recovery outcomes.

Introduction

Problem

Space Y tasks us to train a machine learning model to predict successful Stage 1 recovery!
Analyze different events

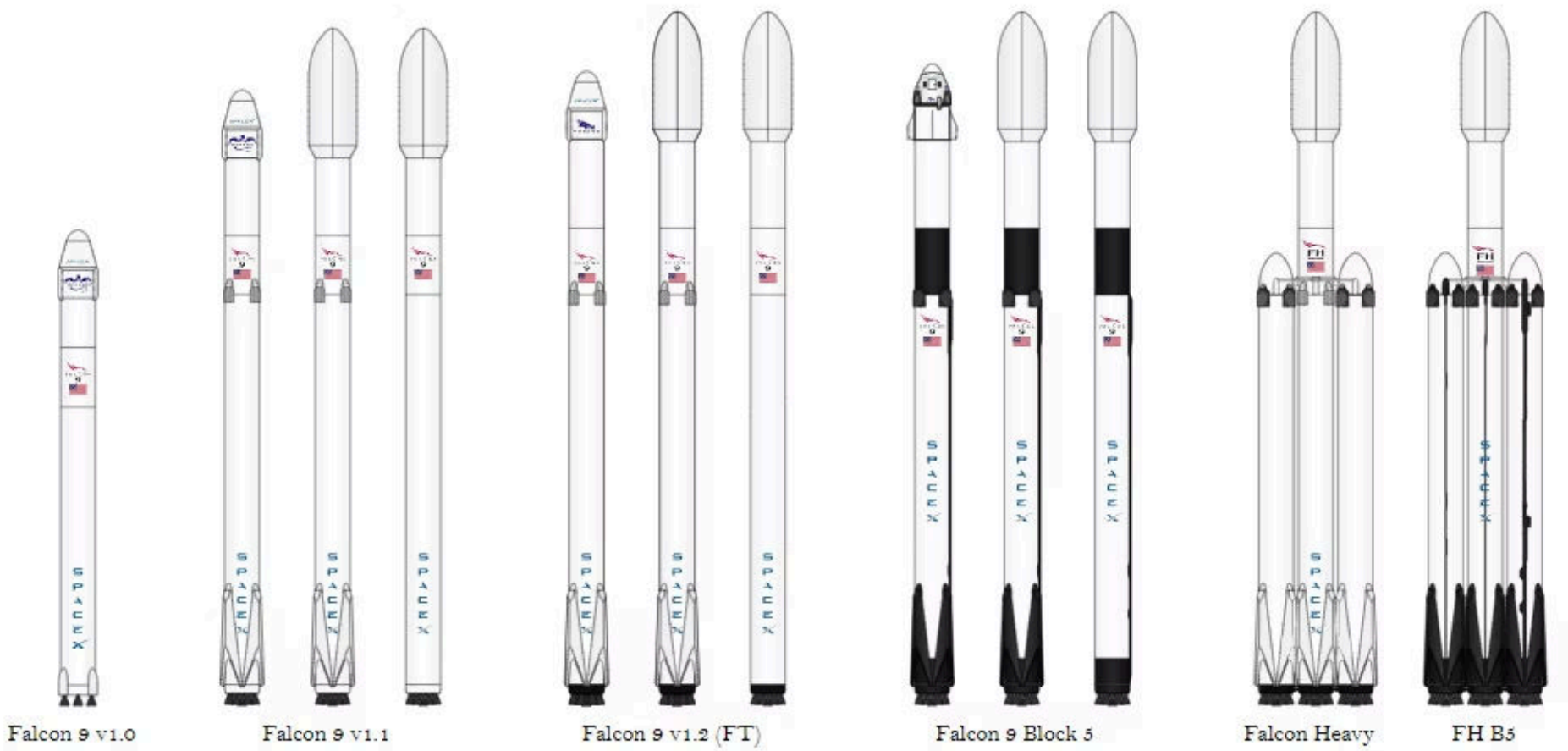
Background

Commercial Space Age is Here !!

- Space X has best pricing ($62 million vs. $165 million USD)
- Largely due to ability to recover part of rocket (Stage 1)
- Space Y wants to compete with Space X

Summary

Collected data from public SpaceX API and SpaceX Wikipedia page. Created labels column ‘class’ which classifies successful landings. Explored data using SQL, visualization, folium maps, and dashboards. Gathered relevant columns to be used as features. Changed all categorical variables to binary using one hot encoding. Standardized data and used GridSearchCV to find best parameters for machine learning models. Visualize accuracy score of all models. • Four machine learning models were produced: Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors. All produced similar results with accuracy rate of about 83.33%. All models over predicted successful landings. More data is needed for better model determination and accuracy.



Methodology Overview

Data Collection

The data was collected from two main sources: SpaceX's public API and their Wikipedia page. The API provided real-time mission data, while Wikipedia offered additional details about past missions. Data was combined to create a comprehensive dataset.

Data Wrangling

The data was cleaned and preprocessed to ensure consistency and remove inconsistencies. Data types were converted, missing values were handled, and redundant information was removed. This step prepared the data for analysis and modeling.

Visualization

Exploratory data analysis (EDA) was conducted using various visualization tools, including SQL and Python libraries. The visualization helped to gain insights into data trends, patterns, and relationships. These insights informed the model selection and feature engineering.

Model Methods

Several classification models were explored to predict the success of Falcon 9 rocket stage 1 recovery. These models were evaluated and compared based on their performance metrics. The best performing model was further tuned using GridSearchCV to optimize its parameters.

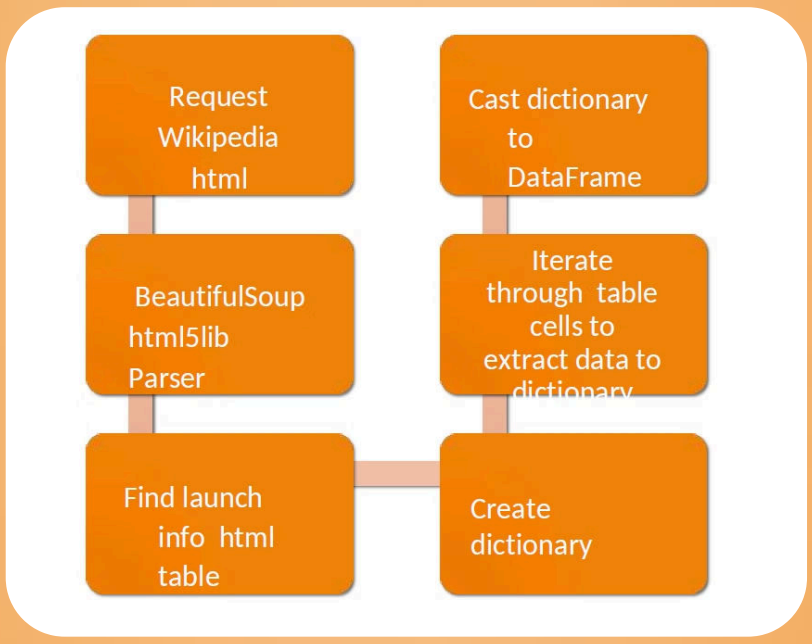
Data Collection

The data collection process involved a combination of API requests from SpaceX's public API and web scraping data from a table on their Wikipedia page. The API provided real-time mission data, while the Wikipedia page offered additional details about past missions.

- The SpaceX API provided data columns such as FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, and Latitude.
- The Wikipedia web scrape provided columns such as Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, and Time.
- The next slide will show a flowchart of the data collection process from the API. The slide after that will show a flowchart of the data collection process from web scraping.

Stages of Data Collection

1. Request (Space X APIs)
2. .JSON file + Lists(Launch Site, Booster Version, Payload Data)
3. Json_normalize to DataFrame data from JSON
4. Filter data to only include Falcon 9 launches
5. Cast dictionary to a DataFrame
6. Dictionary relevant data
7. Imputate missing PayloadMass values with mean



Data Wrangling

Create a training label with landing outcomes where successful = 1 & failure = 0.

Outcome column has two components: ‘Mission Outcome’ ‘Landing Location’

New training label column ‘class’ with a value of 1 if ‘Mission Outcome’ is True and 0 otherwise.

Value Mapping: True ASDS, True RTLS, & True Ocean set to --> 1

None None, False ASDS, None ASDS, False Ocean, False RTLS set to --> 0

Data Visualization

The project uses data visualization to identify trends and insights. Two visualization methods were applied: Folium and Plotly.

Folium is a Python library used to create interactive maps, allowing visualization of launch sites, successful and unsuccessful landings, and the proximity of launch sites to key locations.

1

Folium Maps

Interactive map visualization of launch sites and landing locations, including proximity to key locations.

2

Plotly Dashboard

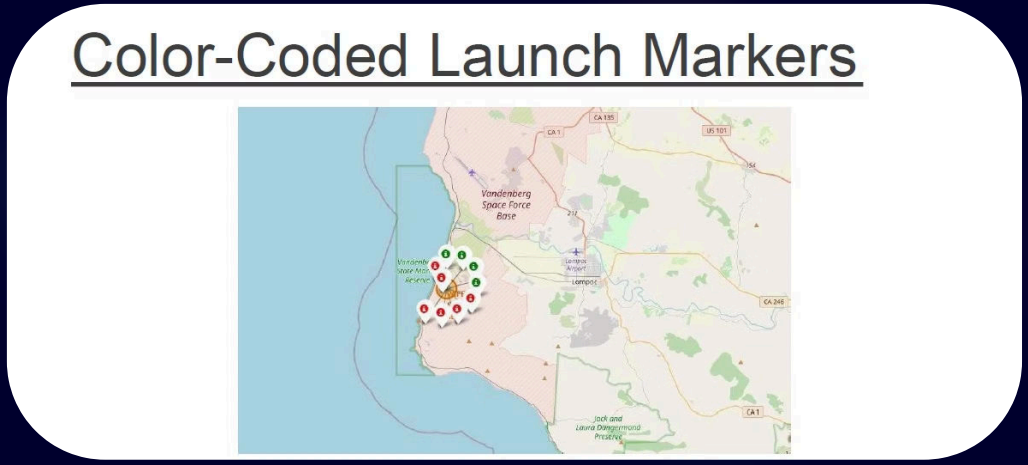
Dashboard with a pie chart and a scatter plot to visualize launch site success rate and relationships between various factors.

3

Interactive Visualizations

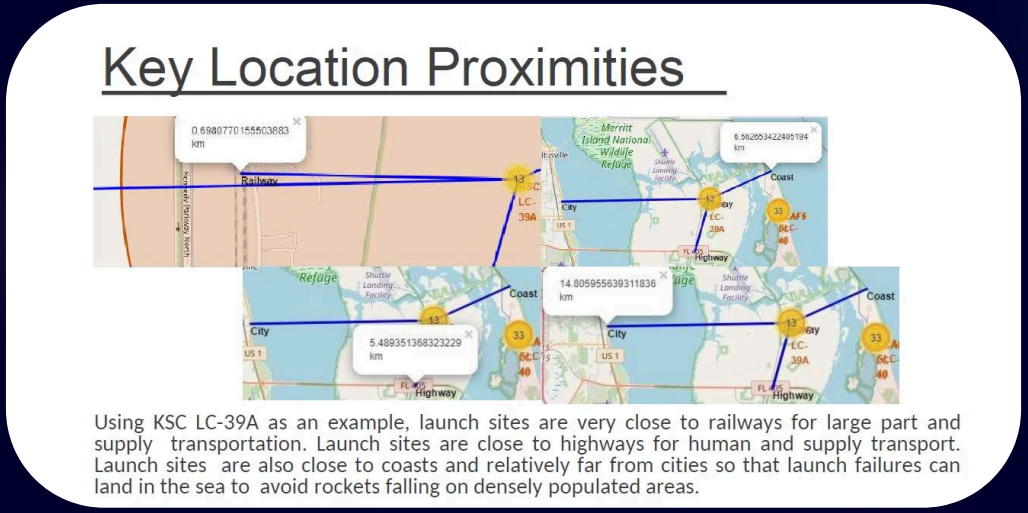
Visualizations provide insights into the factors that influence the success of Falcon 9 rocket landings.

Interactive Map with Folium



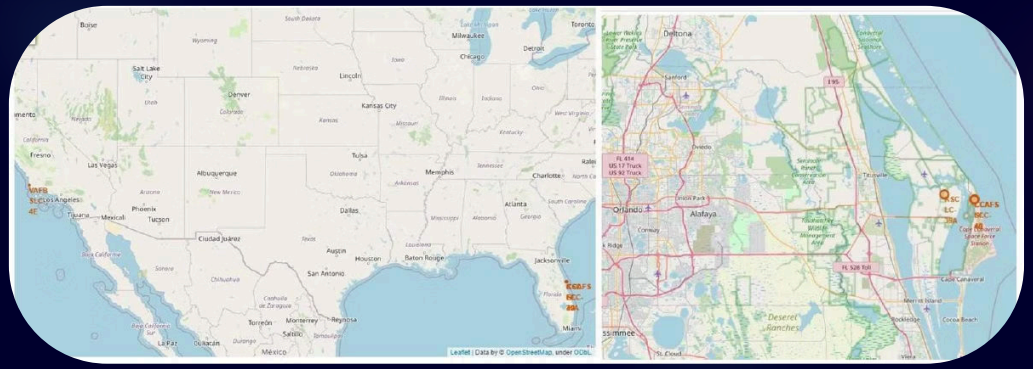
Color Coded Launch Markers

Folium 13 Folium maps mark Launch Sites, successful and unsuccessful landings, and a proximity example to key locations: Railway, Highway, Coast, and City.



Key Location Proximities

This allows us to understand why launch sites may be located where they are. Also visualizes successful landings relative to location.



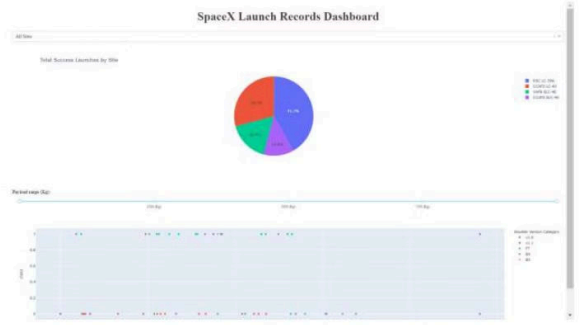
Launch Site Locations

The left map shows all launch sites relative US map. The right map shows the two Florida launch sites since they are very close to each other. All launch sites are near the ocean.

Using KSC LC 39 A as an example, launch sites are very close to railways for large part and supply transportation Launch sites are close to highways for human and supply transport Launch sites are also close to coasts and relatively far from cities so that launch failures can land in the sea to avoid rockets falling on densely populated areas

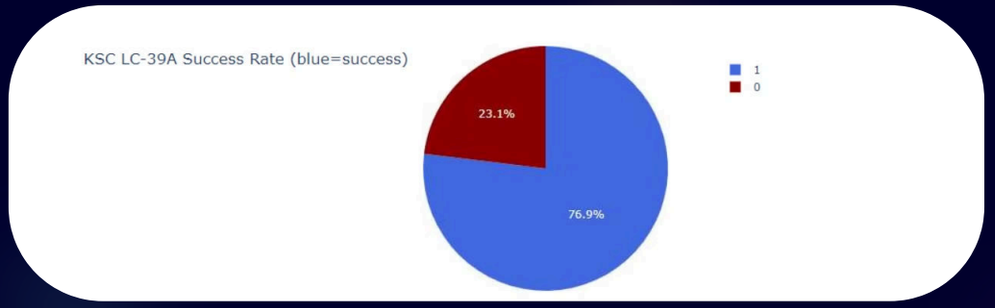
Build a Dashboard with Plotly Dash

Results



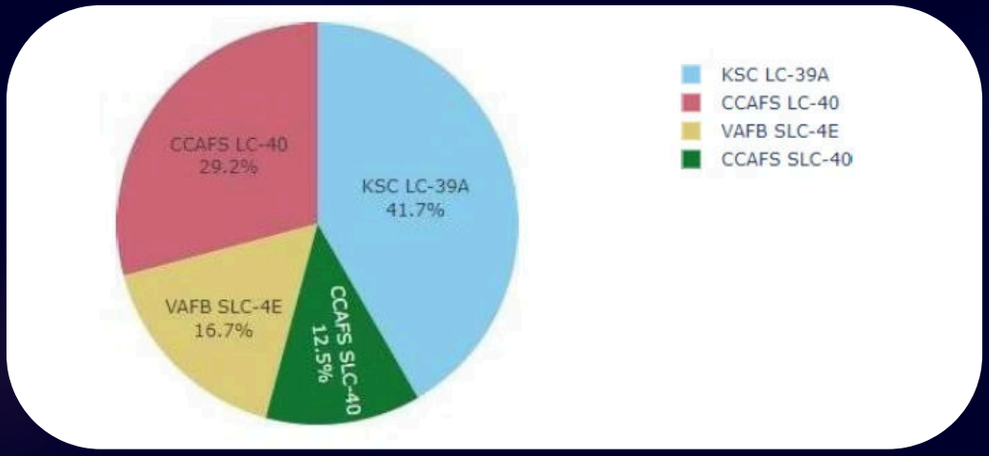
This is a preview of the Plotly dashboard. The following sides will show the results of EDA with visualization, EDA with SQL, Interactive Map with Folium, and finally the results of our model with about 83% accuracy.

Dashboard includes a pie chart and a scatter plot. Pie chart can be selected to show distribution of successful landings across all launch sites and can be selected to show individual launch site success rates. Scatter plot takes two inputs: All sites or individual site and payload mass on a slider between 0 and 10000 kg. The pie chart is used to visualize launch site success rate. The scatter plot can help us see how success varies across launch sites, payload mass, and booster version category.



Highest Success Rate Launch Site

KSC LC 39A has the highest success rate with 10 successful landings and 3 failed landings.



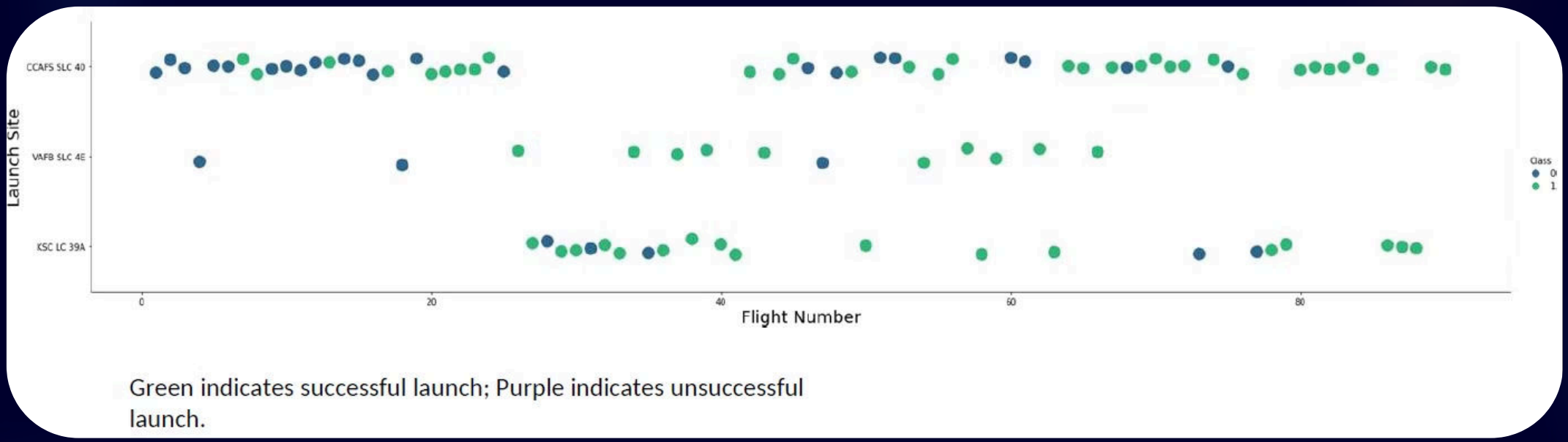
Successful Launches Across Launch Sites

This is the distribution of successful landings across all launch sites. CCAFS LC 40 is the old name of CCAFS SLC 40 so CCAFS and KSC have the same amount of successful landings, but a majority of the successful landings where performed before the name change. VAFB has the smallest share of successful landings. This may be due to smaller sample and increase in difficulty of launching in the west coast.

EDA With Visualization

Exploratory Data Analysis was performed on variables such as Flight Number, Payload Mass, Launch Site, Orbit, Class, and Year. This analysis involved using visualization techniques to explore the relationships between these variables. Scatter plots, line charts, and bar plots were used to compare these relationships and determine if any patterns or trends could be observed.

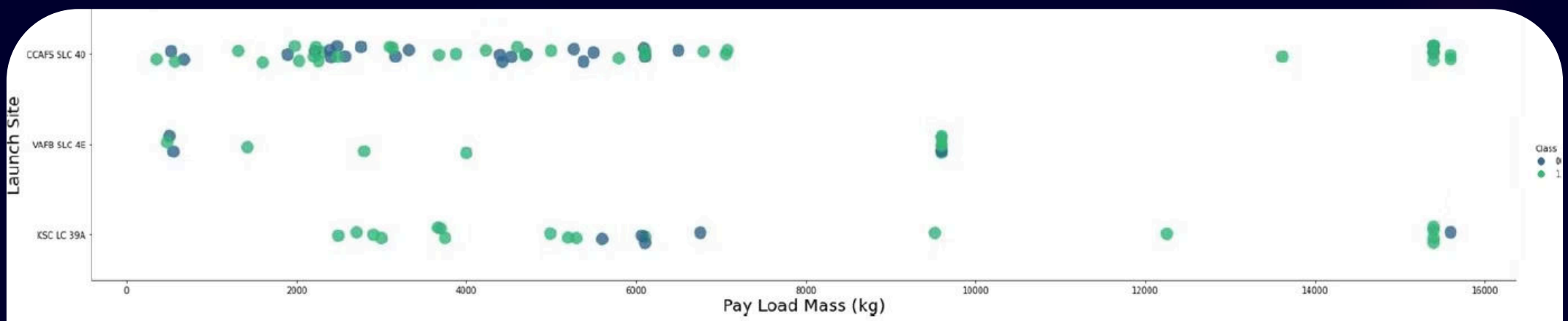
For instance, a scatter plot of Flight Number vs. Payload Mass was used to determine if there was any correlation between the two variables. Similarly, a bar plot of Success Rate vs. Orbit was used to compare the success rate of launches for different orbits. The results of this EDA were used to identify key factors that could influence the success of a Falcon 9 launch.



Green indicates successful launch; Purple indicates unsuccessful launch.

Flight Number vs. Launch Site

Graphic suggests an increase in success rate over time (indicated in Flight Number) Likely a big breakthrough around flight 20 which significantly increased success rate CCAFS appears to be the main launch site as it has the most volume

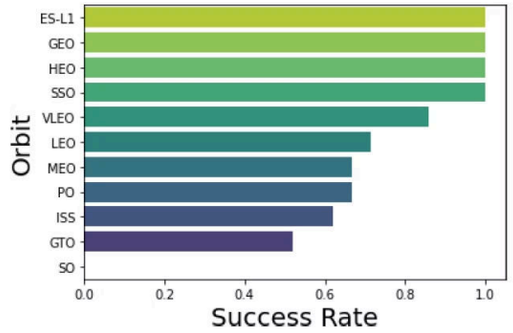


Green indicates successful launch; Purple indicates unsuccessful launch.

Payload vs. Launch Site

Payload mass appears to fall mostly between 0 6000 kg. Different launch sites also seem to use different payload mass.

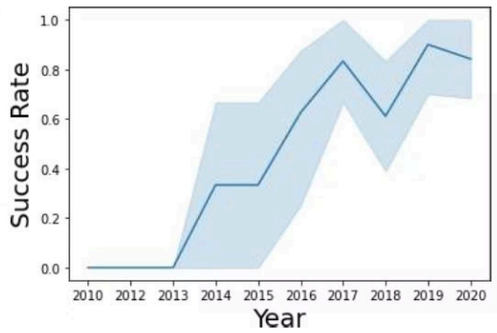
Success rate vs. Orbit type



Success Rate Scale with 0 as 0% 0.6 as 60% 1 as 100%

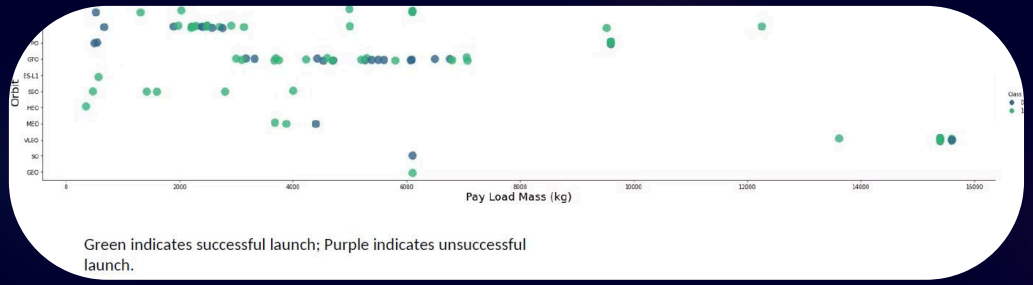
ES L1 (1), GEO (1), HEO (1) have 100% success rate (sample sizes in parenthesis) SSO (5) has 100% success rate VLEO (14) has decent success rate and attempts SO (1) has 0% success rate GTO (27) has the around 50% success rate but largest sample Success Rate Scale with 0 as 0% 0.6 as 60% 1 as 100% 20

Launch Success Yearly Trend



95% confidence interval (light blue shading)

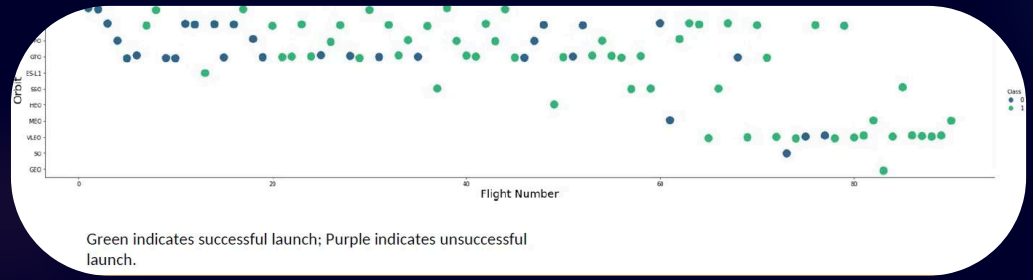
Success generally increases over time since 2013 with a slight dip in 2018 Success in recent years at around 80%



Green indicates successful launch; Purple indicates unsuccessful launch.

Pay load vs Orbit

Payload mass seems to correlate with orbit LEO and SSO seem to have relatively low payload mass The other most successful orbit VLEO only has payload mass values in the higher end of the range



Green indicates successful launch; Purple indicates unsuccessful launch.

Flight Number vs. Orbit type Launch

Launch Orbit preferences changed over Flight Number. Launch Outcome seems to correlate with this preference. SpaceX started with LEO orbits which saw moderate success LEO and returned to VLEO in recent launches SpaceX appears to perform better in lower orbits or Sun synchronous orbits

EDA with SQL

Loaded data set into IBM DB2 Database.

Queried using SQL Python integration.

Queries were made to get a better understanding of the dataset.

Queried information about launch site names, mission outcomes, various pay load sizes of customers and booster versions, and landing outcomes

```
In [4]: %%sql
        SELECT UNIQUE LAUNCH_SITE
        FROM SPACEXDATASET;

         * ibm_db_sa://ftb12020:***@0c77d6f2
        Done.
```

Out[4]:

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

All Launch Site Names : Query unique launch site names from database. CCAFS SLC 40 and CCAFSSLC 40 likely all represent the same launch site with data entry errors. CCAFS LC 40 was the previous name. Likely only 3 unique launch_site values: CCAFS SLC 40, KSC LC 39A, VAFB SLC 4E

```
%%sql
SELECT *
FROM SPACEXDATASET
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[5]:

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

First five entries in database with Launch Site name beginning with CCA.

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS SUM_PAYLOAD_MASS_KG
FROM SPACEXDATASET
WHERE CUSTOMER = 'NASA (CRS)';
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86
Done.
```

sum_payload_mass_kg
45596

This query sums the total payload mass in kg where NASA was the customer. CRS stands for Commercial Resupply Services which indicates that these payloads were sent to the International Space Station (ISS). 27

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD_MASS_KG
FROM SPACEXDATASET
WHERE booster_version = 'F9 v1.1'
```

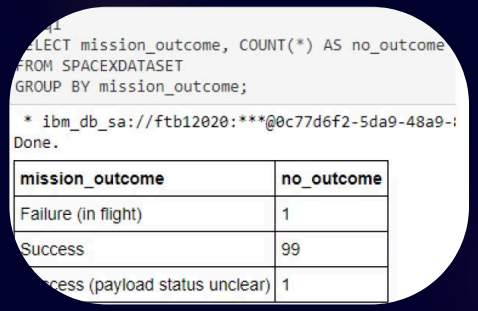
```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-8
Done.
```

avg_payload_mass_kg
2928

This query calculates the average payload mass or launches which used booster version F9 v1.1 Average payload mass of F9 1.1 is on the low end of our payload mass range 28

Successful Drone Ship Landing with Payload Between 4000 and 6000

This query returns the four booster versions that had successful drone ship landings and a payload mass between 4000 and 6000 noninclusively. 30



```
LECT mission_outcome, COUNT(*) AS no_outcome
FROM SPACEXDATASET
GROUP BY mission_outcome;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-
Done.
```

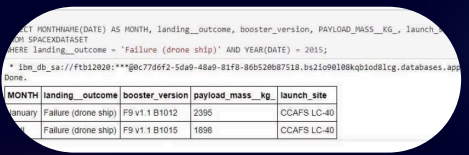
mission_outcome	no_outcome
Failure (in flight)	1
Success	99
cess (payload status unclear)	1

Total Number of Each Mission Outcome

This query returns a count of each mission outcome. SpaceX appears to achieve its mission outcome nearly 99% of the time.

booster_version	payload_mass__kg
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
B5 B1049.7	15600

Boosters that Carried Maximum Payload



MONTH	landing__outcome	booster_version	payload_mass__kg_	launch_site
January	Failure (drone ship)	F9 v1.1 B1012	2395	CCAFS LC-40
	Failure (drone ship)	F9 v1.1 B1015	1898	CCAFS LC-40

2015 Failed Drone Ship Landing Records

This query returns the Month, Landing Outcome, Booster Version, Payload Mass (kg), and Launch site of 2015 launches where stage 1 failed to land on a drone ship. There were two such occurrences.

```
%%sql
SELECT MIN(DATE) AS FIRST_SUCCESS
FROM SPACEXDATASET
WHERE landing__outcome = 'Success (ground pad)';
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81
Done.
```

first_success
2015-12-22

First Successful Ground Pad Landing Date This

```
%%sql
SELECT landing__outcome, COUNT(*) AS no_outcome
FROM SPACEXDATASET
WHERE landing__outcome LIKE 'Succes%' AND DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY landing__outcome
ORDER BY no_outcome DESC;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg
Done.
```

landing__outcome	no_outcome
Success (drone ship)	5
Success (ground pad)	3

Ranking Counts of Successful Landings Between 2010 06 04 and 2017 03 20

Classification

1

Model Selection

The project explored several machine learning models for classification. These included Logistic Regression, K-Nearest Neighbors, Decision Tree, Random Forest, Support Vector Machine (SVM), and Naive Bayes.

2

Model Evaluation

Each model's performance was evaluated using metrics like accuracy, precision, recall, and F1-score. These metrics helped assess the model's ability to correctly classify successful and unsuccessful landings.

3

Best Performing Model

After comparing the performance of different models, the Random Forest Classifier emerged as the best performing model. It achieved the highest accuracy, precision, recall, and F1-score on the test set.

4

Model Tuning

The Random Forest model was further tuned using GridSearchCV to optimize its hyperparameters. This process involved systematically searching for the best combination of hyperparameters to maximize the model's performance.

Classification

1

Model Comparison

All models performed similarly with an accuracy of 83.33% on the test set. The test set had a small sample size of 18, which can lead to large variance in accuracy results.

2

Confusion Matrix

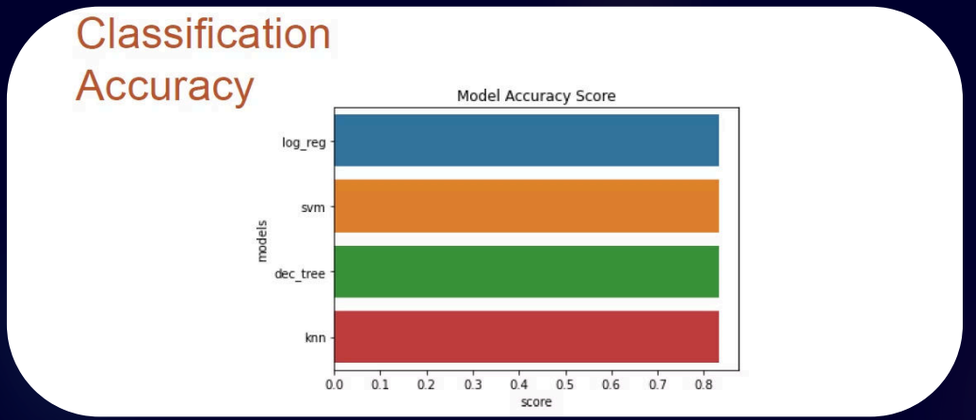
The confusion matrix shows that all models have the same performance due to the small test size. The models predicted 12 successful landings correctly. They overpredicted 3 successful landings when the true label was unsuccessful landing.

3

Data Needs

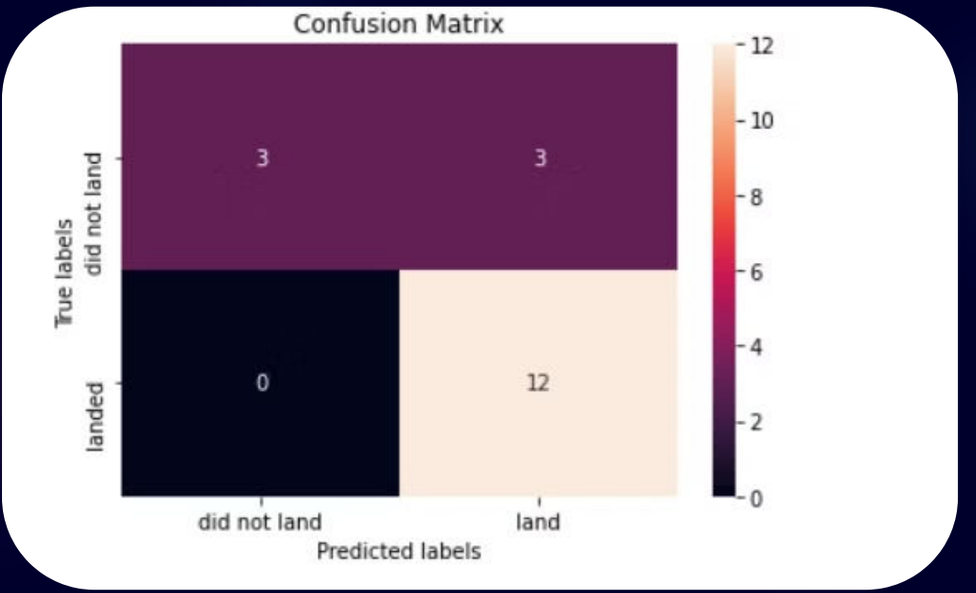
To determine the best model, more data is required. A larger test set will give more statistically significant results and help identify the model that performs best in predicting landing success.

Predictive Analysis



Classification accuracy

The classification accuracy is a key metric in evaluating the performance of our machine learning model. By thoroughly analyzing the model's classification accuracy, we can identify areas for improvement and fine-tune the model to achieve optimal results.



Confusion matrix

Since all models performed the same for the test set, the confusion matrix is the same across all models. The models predicted 12 successful landings when the true label was successful landing. The models predicted 3 unsuccessful landings when the true label was unsuccessful landing. The models predicted 3 successful landings when the true label was unsuccessful landings (false positives). Our models over predict successful landings.

Conclusion

Our project aimed to develop a machine learning model for Space Y, a company seeking to bid against SpaceX. The model predicts successful Stage 1 landings, saving Space Y millions of dollars. This model was trained using data collected from the public SpaceX API and web scraping SpaceX's Wikipedia page.

We built a dashboard for visualization, and our model achieved an 83% accuracy. This enables Space Y to predict with reasonable accuracy whether a launch will have a successful Stage 1 landing before launch, informing their decision to proceed. However, more data collection is crucial to determine the optimal machine learning model and enhance accuracy.

Thank You

- Our task: to develop a machine learning model for Space Y who wants to bid against SpaceX
- The goal of model is to predict when Stage 1 will successfully land to save ~$100 million USD
- Used data from a public SpaceX API and web scraping SpaceX Wikipedia page
- Created data labels and stored data into a DB2 SQL database
- Created a dashboard for visualization
- We created a machine learning model with an accuracy of 83%