Capstone, IBM Data Science Series

SpaceX Launch Success

Executive Summary

- In this project, we analyze the success rate of SpaceX Falcon 9 launches.
- We demonstrate that the success rate has improved over time.
- We use several machine learning algorithms and predictive analytic tools.

Introduction

- SpaceX advertises a Falcon 9 launch cost of $62 million, in comparison to a cost of up to $165 million from other providers.
- Reusability is a major factor driving SpaceX's cost advantage.
- Being able to predict the success of a SpaceX launch is important for SpaceX's business and for a potential competitor.

Data Collection

- Launch data is available from SpaceX's API: https://api.spacexdata.com/v4/launches/past
- For this project, we are using a stored dataset hosted separately.
- After parsing into a dataframe, the data contains 187 rows and 43 columns. After deleting some columns and filtering to Falcon 9, there are 90 rows and 17 columns.

Data Wrangling

- The largest share of launches are to Geostationary Transfer Orbit (GTO).
- Landing outcomes are tracked in the 'Class' column, with a 0 representing a failed launch and 1 a success. About 67% of the entries in our data set were successful.

```
# Apply value_counts on Orbit column
df["Orbit"].value_counts()
```

```
GTO      27
ISS      21
VLEO     14
PO        9
LEO       7
SSO       5
MEO       3
ES-L1     1
HEO       1
SO        1
GEO       1
Name: Orbit, dtype: int64
```

```
df["Class"].mean()
```

```
0.6666666666666666
```

EDA and Interactive Visual Analytics

Visual analytics are crucial for understanding a data set.

We demonstrate the following visualizations:

- Information about the dataset with Seaborn,
- Maps with Folium,
- An interactive dashboard with Plotly.

Predictive Analysis Methodology

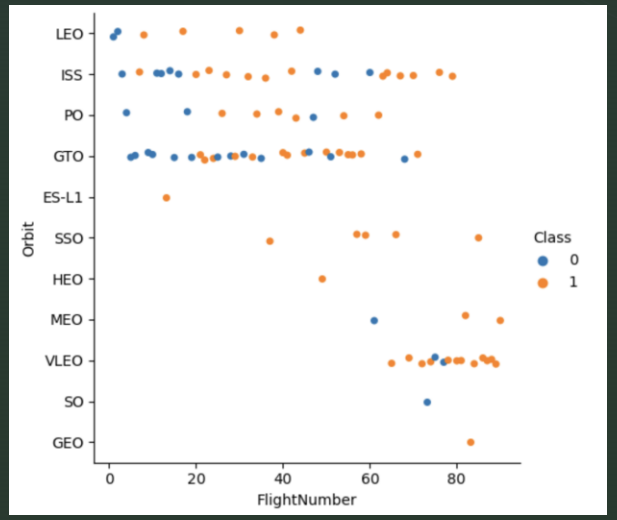
Our analysis demonstrates the following techniques:

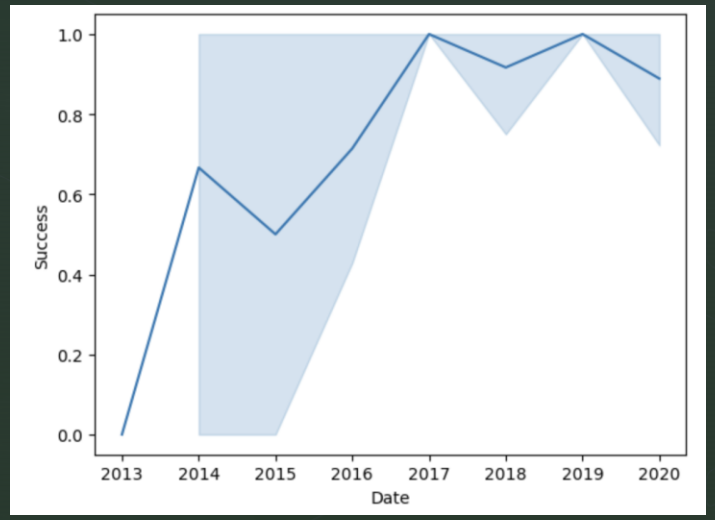
- Exploratory data analysis with SQL,
- Four machine learning algorithms,
- Train-test split,
- Model evaluation with accuracy and a confusion matrix.

EDA Visualization

Using Seaborn to Visualize the Data

- Graphical tools such as Seaborn provide much useful information about the data set.
- Left: successful (Class 1) and unsuccessful (Class 0) flights by orbit and flight number. Right: the data shows increasing success rates over time.





EDA and SQL

- The earliest successful landing was on December 22, 2015.
- There have been several landing outcomes from June 4, 2010 to March 20, 2017.

```
%sql SELECT Min(Date) AS "Earliest Successful Landing" FROM SPACEXTABLE WHERE Landing_Outcome = "Success (ground pad)";
```

```
 * sqlite:///my_data1.db
Done.
```

Earliest Successful Landing
2015-12-22

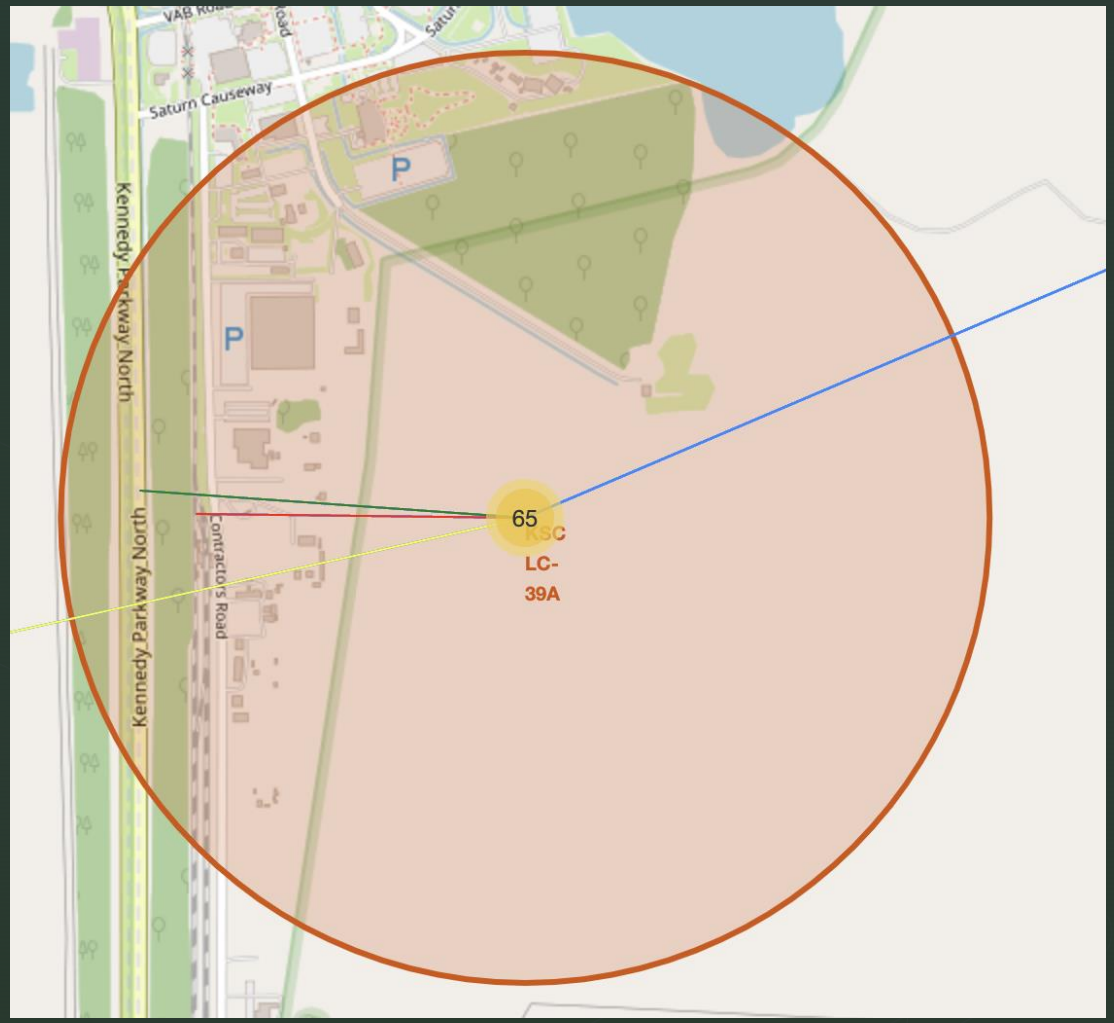
```
%sql SELECT Landing_Outcome AS "Outcome",COUNT(*) AS C FROM SPACEXTABLE \
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' \
GROUP BY Landing_Outcome \
ORDER BY C DESC;
```

```
 * sqlite:///my_data1.db
Done.
```

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

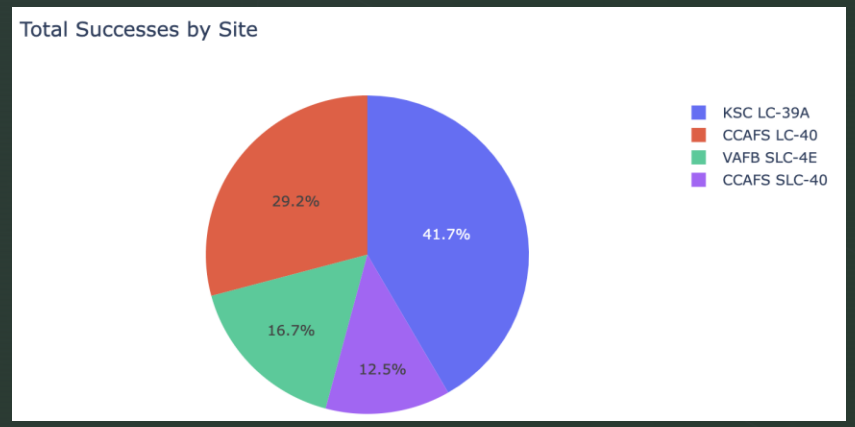
Folium

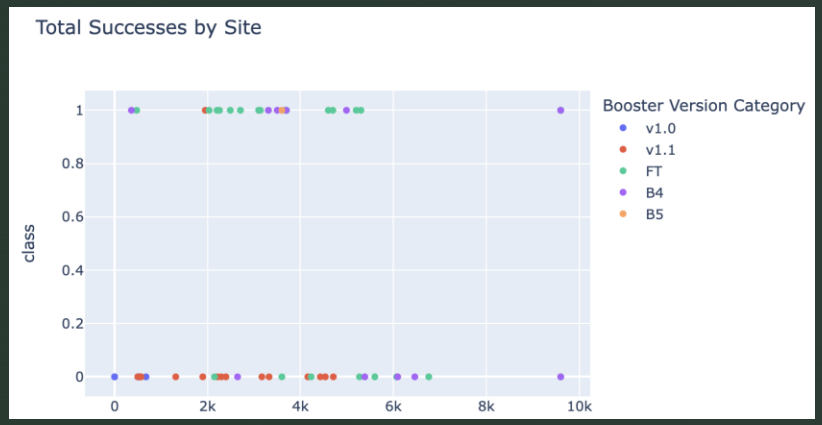
- Folium creates interactive maps that yield useful insights about the data.
- Shown below is that a launch site, the Kennedy Space Center, is located near a railroad (red line) and a highway (green). All centers are close to the coast (blue) and at Southern latitudes. Due to noise and risks from rockets, the centers cannot be too close to towns (yellow).

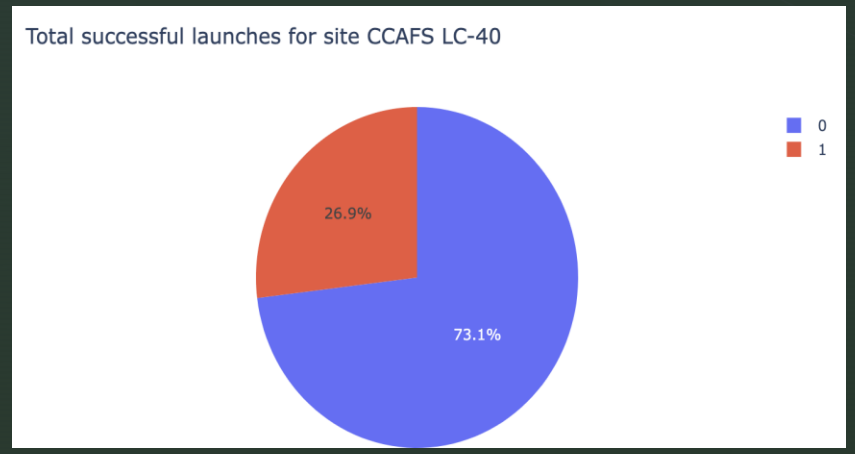


Plotly

- Plotly is a python-based tool for creating interactive visualizations.

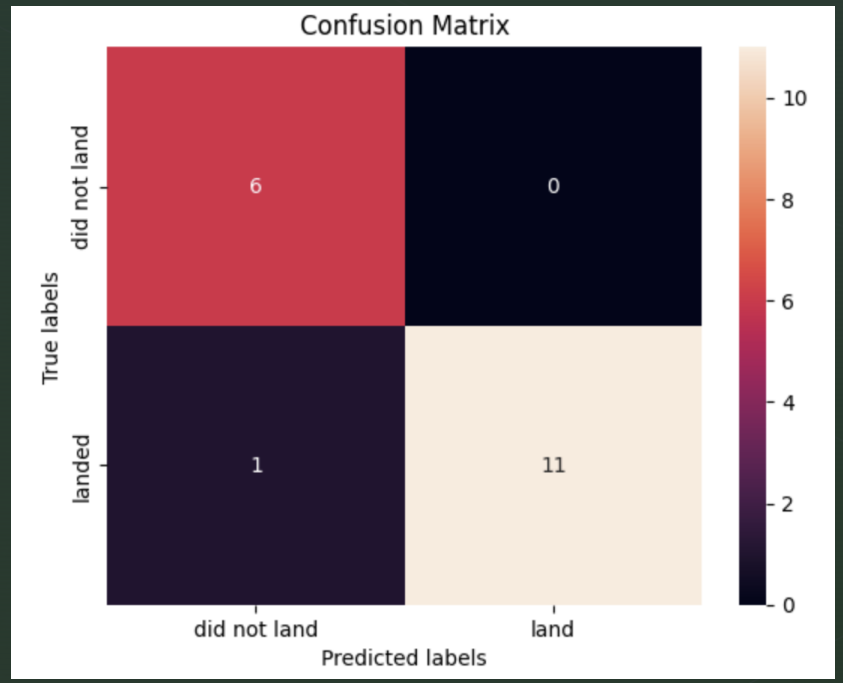






Predictive Analysis

- We tried four learning methods: Logistic Regression, Support Vector Machines, Decision Tree, and K-Nearest Neighbors. We found that the decision tree performed yet.
- The limited sample size (90 samples, 18 test samples) make it difficult to be too confident in these results.



Conclusions

- SpaceX's launch success rate has improved over time.
- Though there is too little data to be confident, it appears from the superiority of support vector machines and decision trees over logistic regression and k-nearest neighbors that launch success is a nonlinear phenomenon in the input data.
- There is no visually obvious relationship between flight success and intended orbit, when mediated by date.