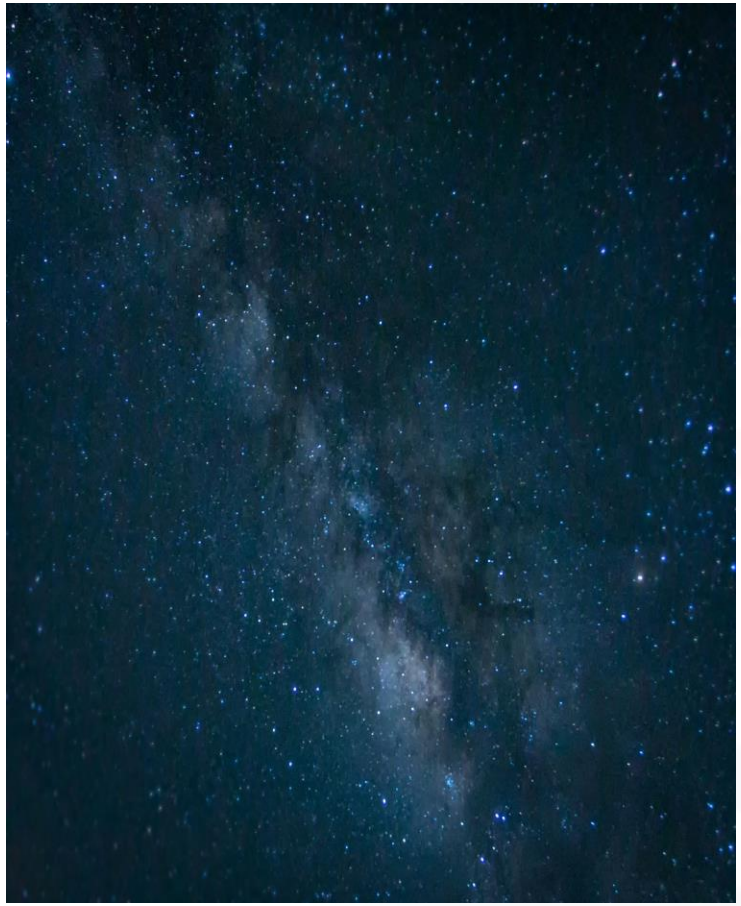
SPACEX PROJECT

Saidat Ibiribigbe

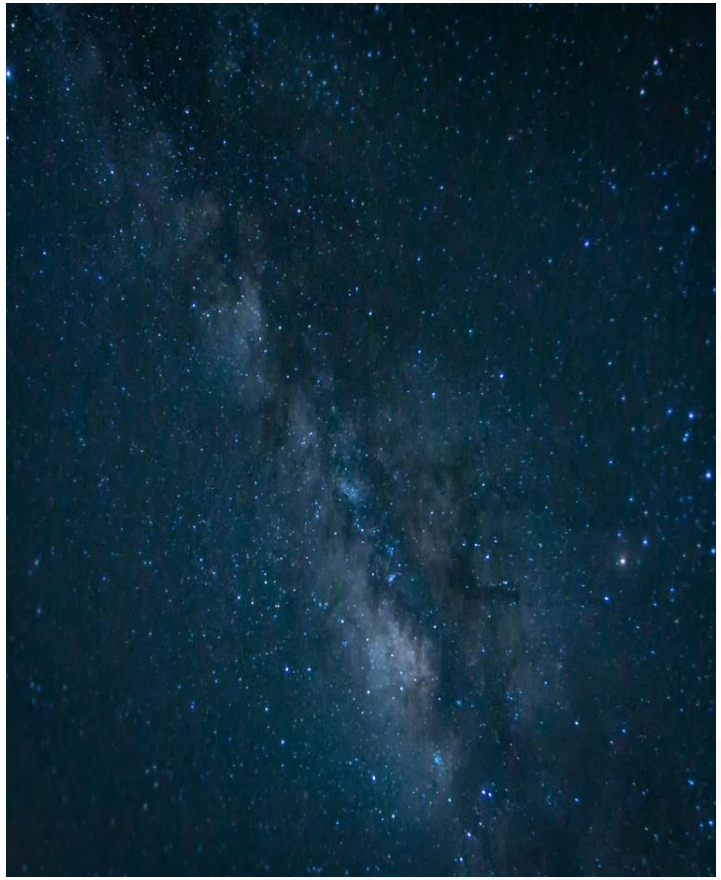
July 16, 2022

Outline



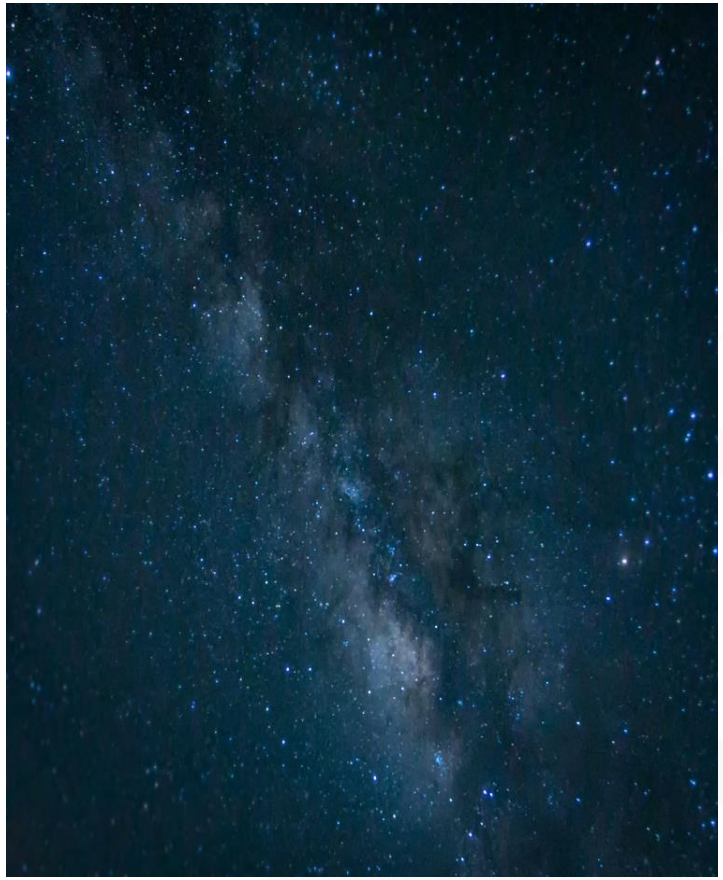
- Summary
- Introduction
- Data Collection
- Data Analysis
- Exploration Analysis
- Methodology
  - Predictive Models
  - Model Selection Results
- Conclusion

Summary



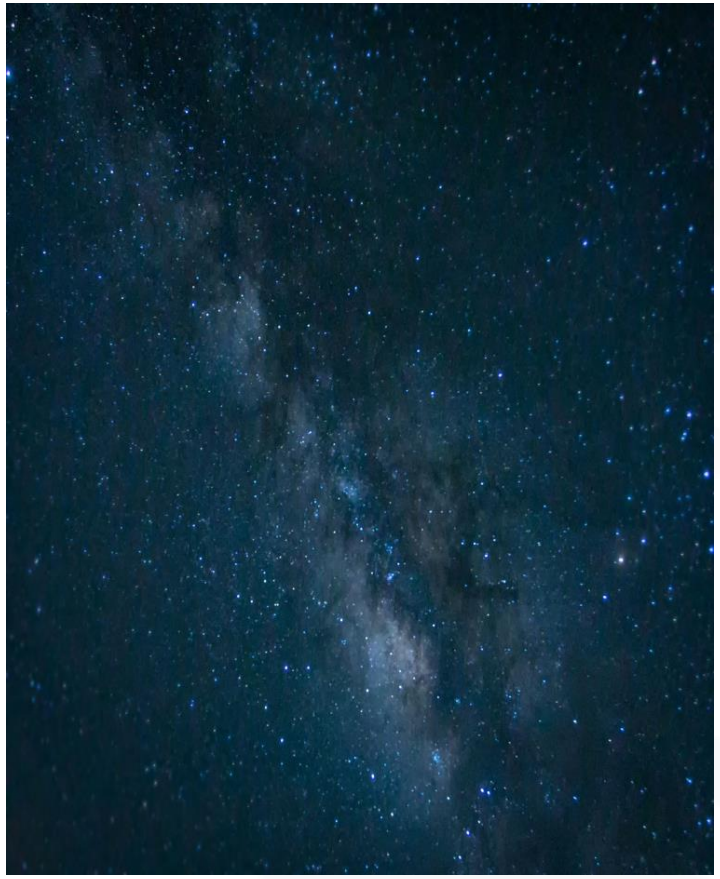
- The purpose of this project is to explore SpaceX data to predict whether a Falcon 9 launch will land successfully and the costs associated with each landing.
- Exploratory data analysis was performed using various tools. Data visualizations were created to analyze trends across launch sites. Geospatial visualizations were also used to draw conclusions on the locations of various launch sites.
- For predictive modeling, four supervised learning algorithms were explored and the best algorithm was suggested based on its predictive power.

Introduction



- To better understand the different factors that influence Falcon 9 launches and identify potentially important factors, exploratory data analysis was performed on the data using various visualization and query techniques.
- The information gained from the data exploration was used to explore various modeling algorithms to determine the best model that presents the most accurate prediction for the proposed business problem.
- Drawing insight from modeling results, potential process improvements and alternate methods are proposed at the end of the presentation.

Data Collection and Wrangling



- Launch data was scraped from a website URL and reformatted into a data frame for easy analysis in Python.
- The scraped data was further cleaned and filtered to only include Falcon 9 launches and variables specific to our analysis requirements.
- Some variables, e.g. payload mass, had missing values that were imputed with the mean of the column as a best estimate.

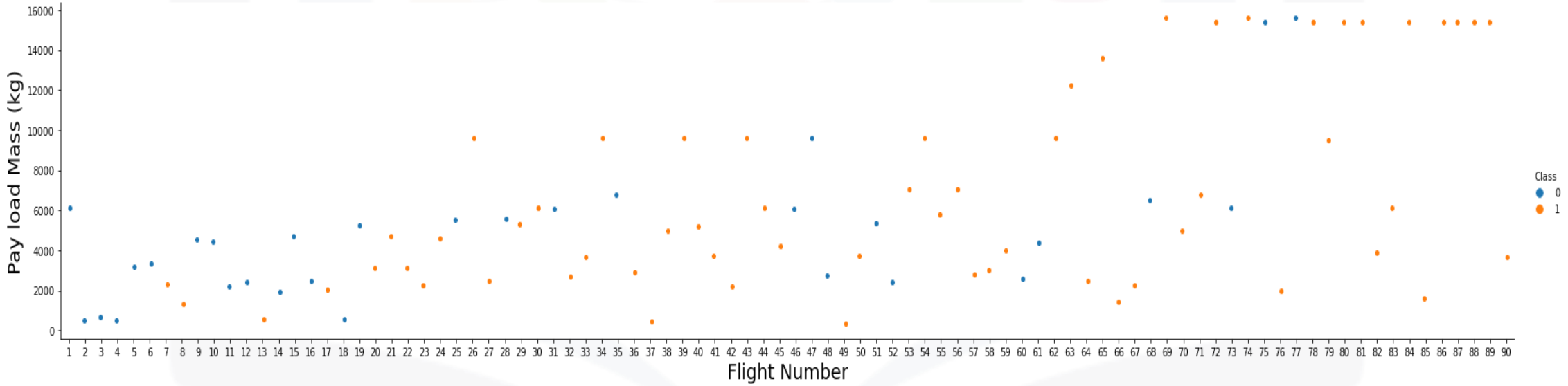
Data Analysis - SQL

Our data analysis begins with query that provided interesting insights. We find that the there are four unique launch sites: CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E. The first successful landing outcome was recorded in July 2018. Additionally in 2015, the only launch failures that occurred happened at the CCAFS LC-40 launch site. Finally, the landing outcome and their corresponding occurrences between June 4, 2010 and March 20, 2017 are shown below. We see that a landing outcome has a higher probability of being recorded as not attempted.

landing__outcome	numofoccurence
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

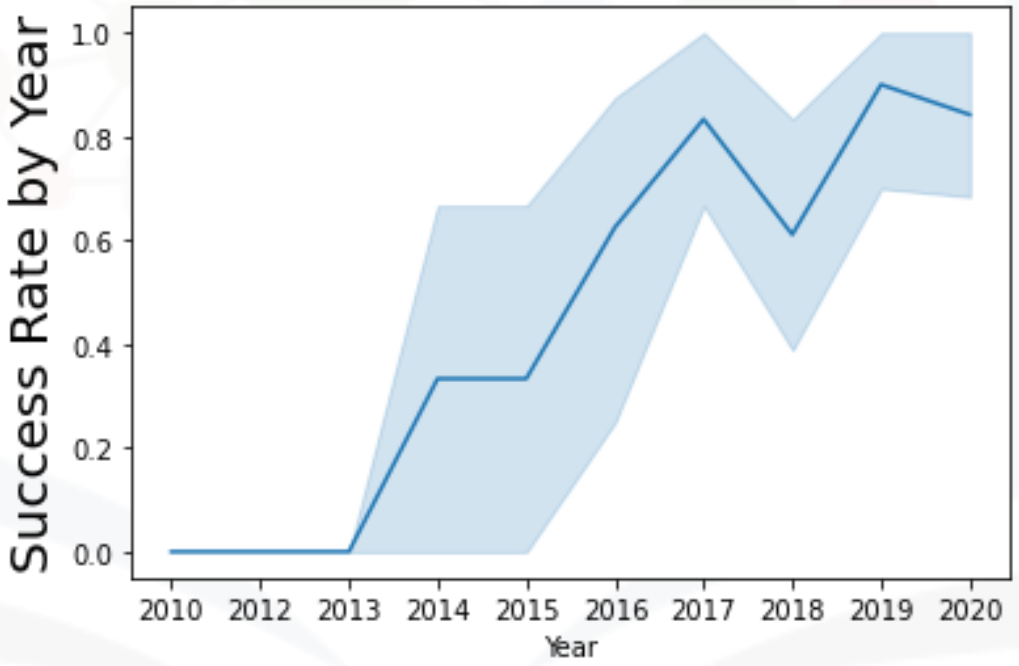
Data Analysis – Success Rate by Flight Number

A deeper analysis was done using visualizations created with seaborn and pyplot. We see that later launches have higher payload mass in kilograms but are also more likely to be successful than earlier launches. This could be due to be the fact that lessons learned in earlier failed launches are applied in later launches to improve the probability of success.



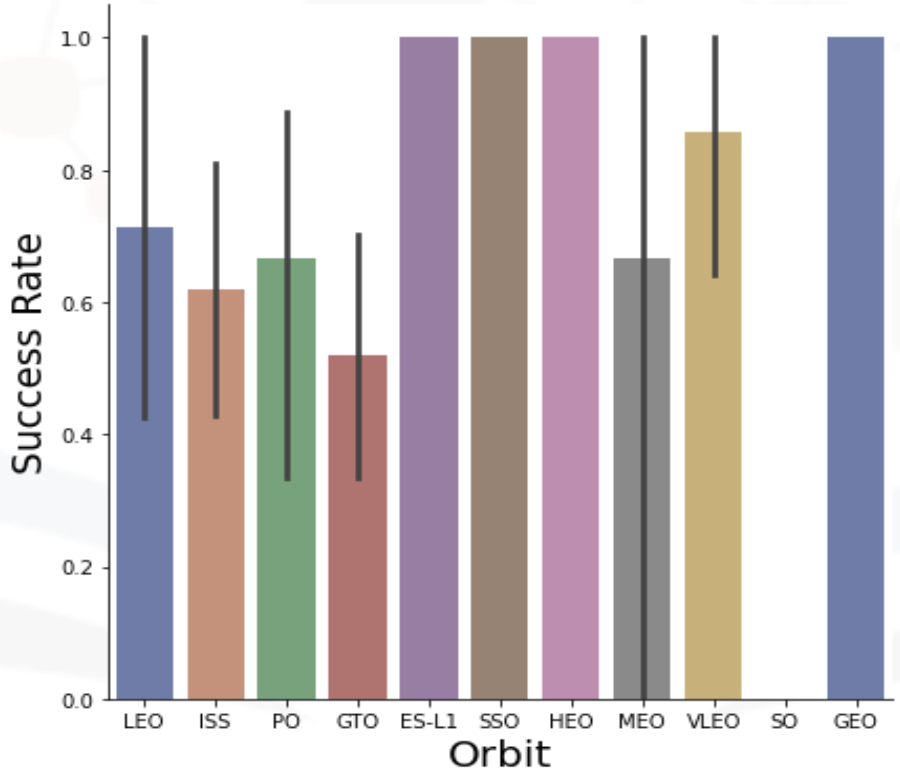
Data Analysis – Success Rate by Year

A line plot of the success rate by year shows a steady increase in launch success rate between 2013 with the exception of a 20% decrease in success rate in from 2018 to 2019.



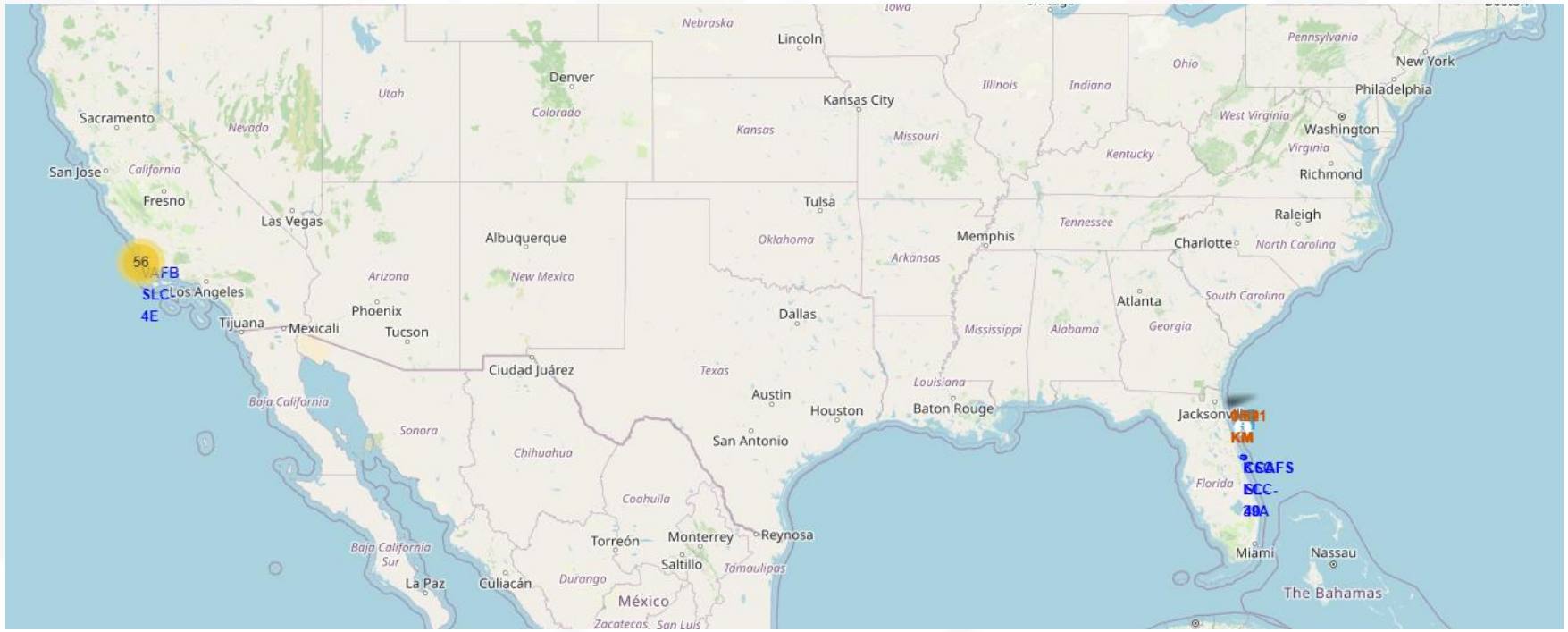
Data Analysis – Success Rate by Orbit

As seen in the chart below, four orbits – GEO, HEO, SSO, ES-L1, all have a 100% success rate. This could be due to the fact that there are fewer launches in these orbits or launches in the orbit occurred at a later time, increasing the chances of success.



Exploratory Analysis – Site Geography

Visual data exploration was also done by creating maps using Folium. Geolocation analysis showed that launch sites are more likely to be closer to coastlines than cities. The VAFB launch site is on the West coast of the United States, while the remaining 3 launch sites are on the East Coast.

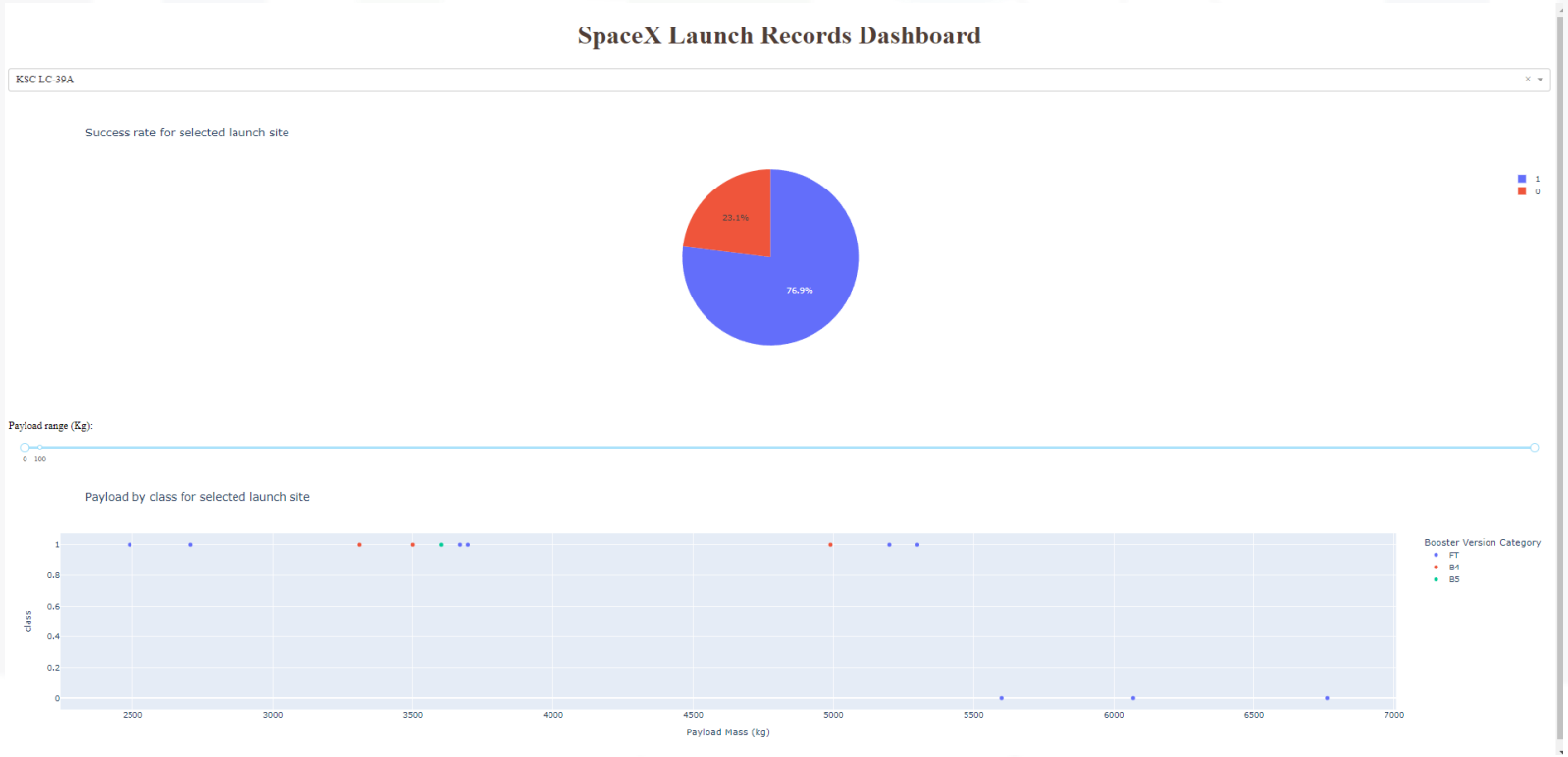


Exploratory Analysis - Interactive Visualizations

An interactive dashboard was created using Dash that provided some insights into launch sites and their success rate with regards to payload. For all launch sites, the payload ranged between 0 and 9,600 kg with success being more likely with lower payloads than higher payloads. We also observed that launch site KSC LC-39A was the most used launch site.

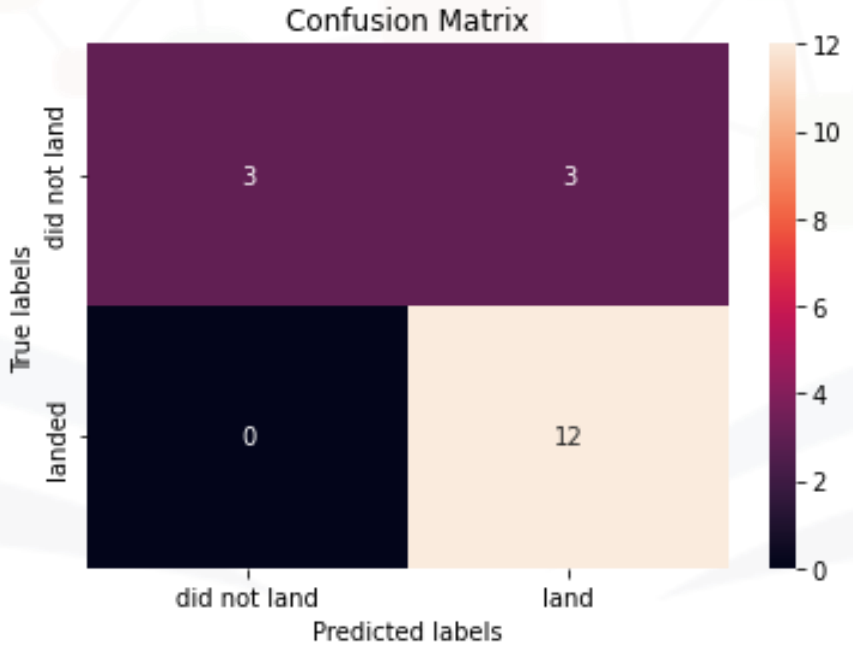
Exploratory Analysis - Interactive Visualizations

Upon further exploration, we observe that launch site also KSC LC-39A has the highest success rate (77%) with success more likely in payloads mass between 2500 and 5500 kg. This helps explains why it is used more frequently, since launches from this site appear to have a higher chance of success.



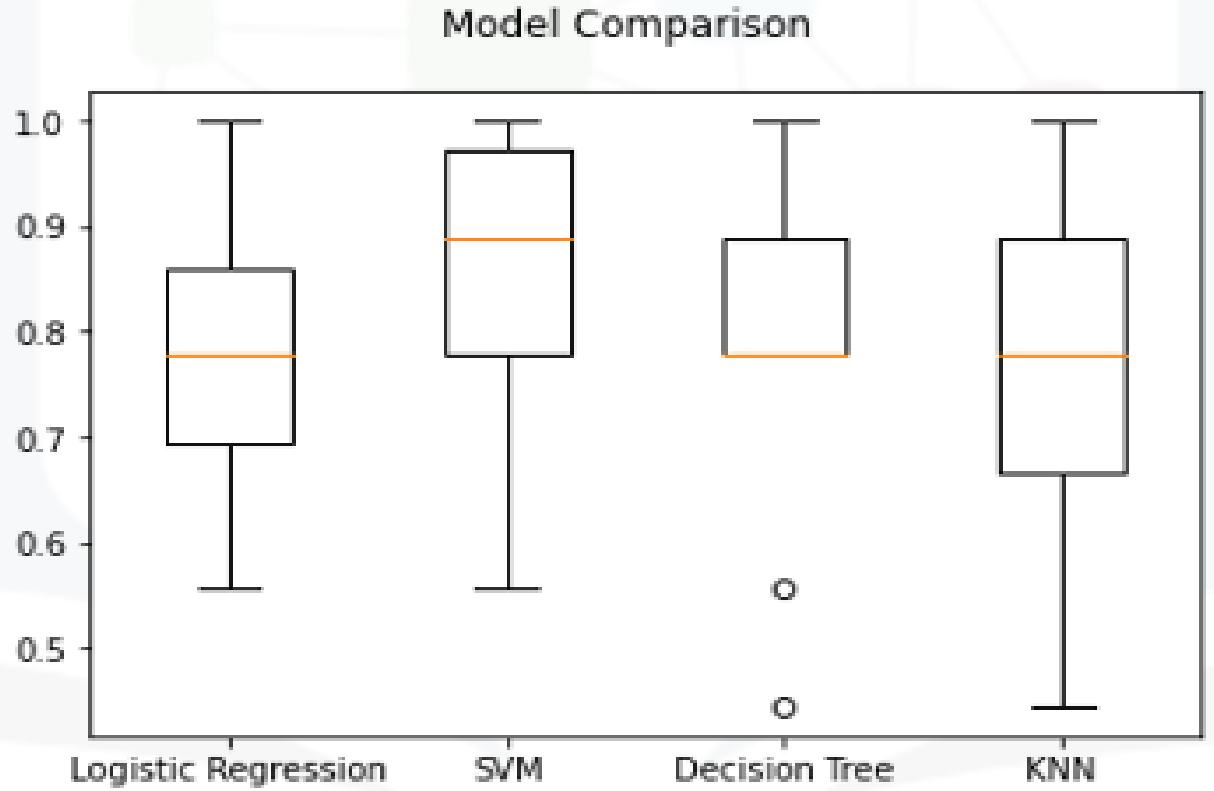
Methodology – Predictive Modeling

Prior to modeling, the dataset was standardized and preprocessed to minimize bias and improve precision. Four supervised learning algorithms were explored using a penalty that helps improve the predictive accuracy in our models, since dataset is not robust. The data was split using the 80/20 ratio, which created a test set of 18 observations. These observations were used to test the predictive performance of the algorithms. For each algorithm, a confusion matrix was constructed on previously unseen data to help assess model performance. The support vector machine model has the highest test accuracy with a value of 89%. It's confusion matrix is displayed below.

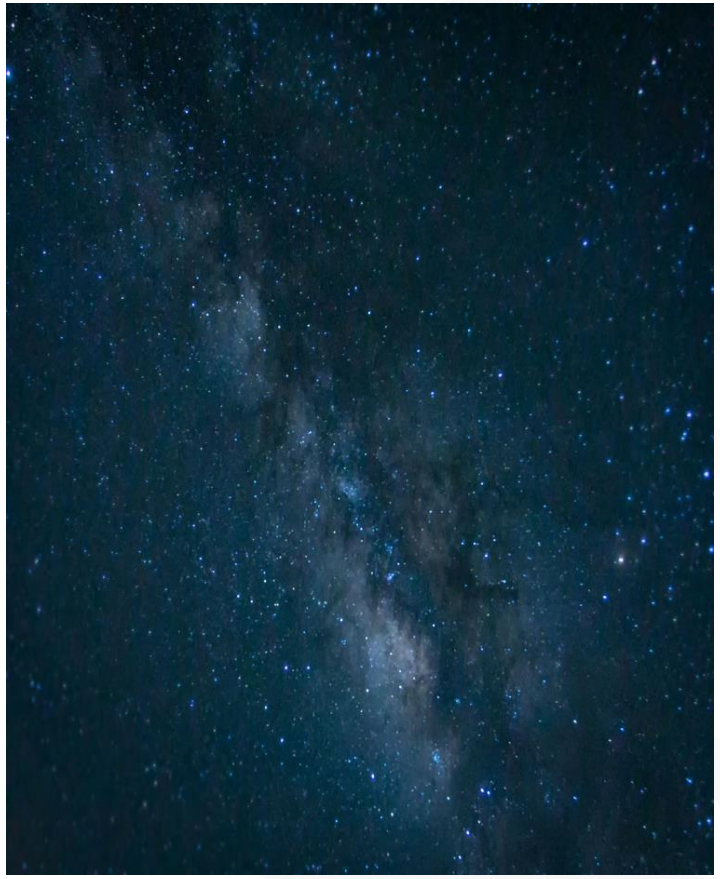


Methodology - Results

To identify the best algorithm, the plot below was created. It shows the distribution of all four algorithms' prediction accuracy scores. From the output we can see that the support vector algorithm and logistic regression algorithm perform better than the decision tree and KNN classifier.



CONCLUSION



- From the outcome of the data analysis, we can say that the most important factors that influence the success of a launch include but are not limited to: launch site, payload mass and launch year.
- With this information, for future launch predictions, a support vector machine model should be used. This is because its high test accuracy score 89% provides us with the best chance of accurately predicting launch outcome.
- Since we only explored supervised learning models for this study, an improvement in prediction accuracy can be achieved by exploring more robust techniques like random forests or gradient boosting. We can also explore unsupervised learning algorithms to help find more hidden and actionable insights and relationship in our data.