IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Michael Bonacci

June 5, 2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection using web scraping and SpaceX REST API calls;
  - Exploratory Data Analysis (EDA), including data wrangling, data visualization and interactive visual analytics;
  - Machine Learning Prediction
- Summary of all results
  - It was possible to collect valuable data from public sources;
  - EDA allowed to identify which feature are the best to predict succes of launchigs;
  - Machine Learning Prediction showed the best model to predict which characteristics are important to drive this opportunity by the bbest way, using all collected data.

Introduction

Nature of Analysis

- SpaceX revolutionized the aerospace industry with the Falcon 9's cost efficiency, driven by it's first stage's reusability.
- Falcon 9's first stage can land back on Earth after launch, ready to be reused for another mission.
- This capability significantly lowers the cost of space travel, becoming an industry game-changer.
- The objective is to evaluate the viability of SpaceY's ability to effectively compete with SpaceX.

Challenge Statements

- The use of machine learning models to predict whether the first stage of the Falcon9 rocket will successfully land, based on data collected from previous launches.
- Estimate the total costs for launches.

Objective Analysis Statements

- Exploratory Data Analysis (EDA): Analyzing & undestanding the dataset to identify patterns, trends and key features related to the landing outcome.
- Data Preparation: Preprocessing the data, standardizing it & splitting the data into training and testing sets.
- Model Building: Training & evaluating multiple machine learning models, including but not limited to Logistic Regression, Decision Trees, Support Vector Machines, K-Nearest Neighbors.
- Model Evaluation: Asessing the models' performance choosing the best method for accuracy.
- Conclusions: Summarizing the key findings & selecting the best performing model.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX REST API calls
  - Web scraping Wikipedia's "List of Falcon 9 & Falcon Heavy launches' using BeautifulSoup.
- Perform data wrangling
  - Convert data into a more usable form
  - Determine Training Labels (1 = the booser successfully landed; 0 = it was unsuccessful)
  - Prepared collected data with use of feature engineering which created landing outcome label based on analyzing & summarizing outcome data.
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Understand the SpaceX Dataset
  - Load the dataset into the corresponding table in a Db2 database.
  - Execute SQL queries to solve questions for SpaceX/Y project.

Data Collection

- SpaceX REST API calls (https://api.spacexdata.com/v4/rockets/launches/past) using Wikipedia
- Web scraping Wikipedia's "List of Falcon 9 & Falcon Heavy launches' using Beautiful Soup
  https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Methodology

Executive Summary

Perform interactive visual analytics using Folium and Plotly Dash

- Mark all launch sites using Folium maps
- Mark the success/failed launches for each site on the map
- Calculate the distances between a launch site to its proximities.

Perform predictive analysis using classification models

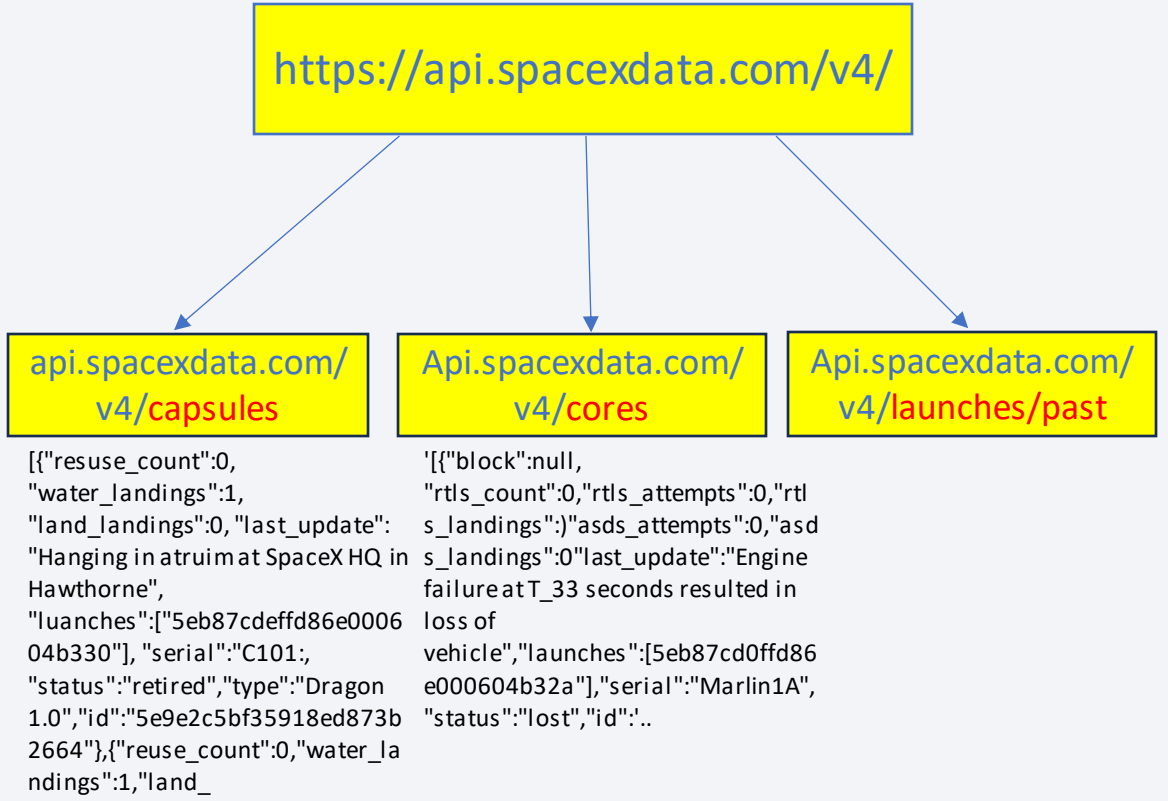
- Perform Exploratory Data Analysis & determine Training Labels
- Create a column for the class
- Standardize the data
- Split into training data and test data
- Find the best Hyperparameter for SVM, Classification Trees and Logistic Regression.

Data Collection – SpaceX API

- SpaceX REST API calls obtained using the following link:

https://api.spacexdata.com/v4/launches/past

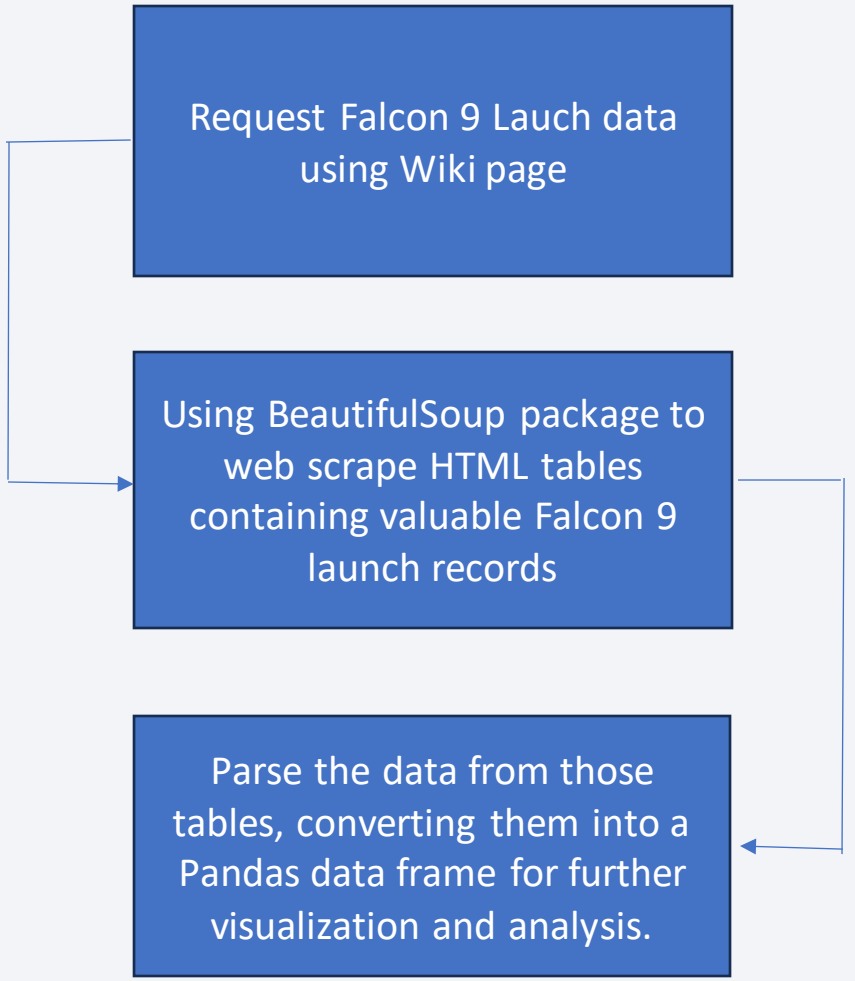
- The API was executed in accordance with the flowchart in example 1 (to the right).
- https://github.com/gotnerd/Space-X-Falcon-9-First-Stage-Landing-Prediction-Lab-2-Data-wrangling/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



Data Collection - Scraping

- Data Collection use of SpaceX launches obtained using Wikipedia;
- Data from the flowchart can be obtained using the following link:

- https://github.com/gotnerd/Space-X-Falcon-9-First-Stage-Landing-Prediction-Lab-2-Data-wrangling/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



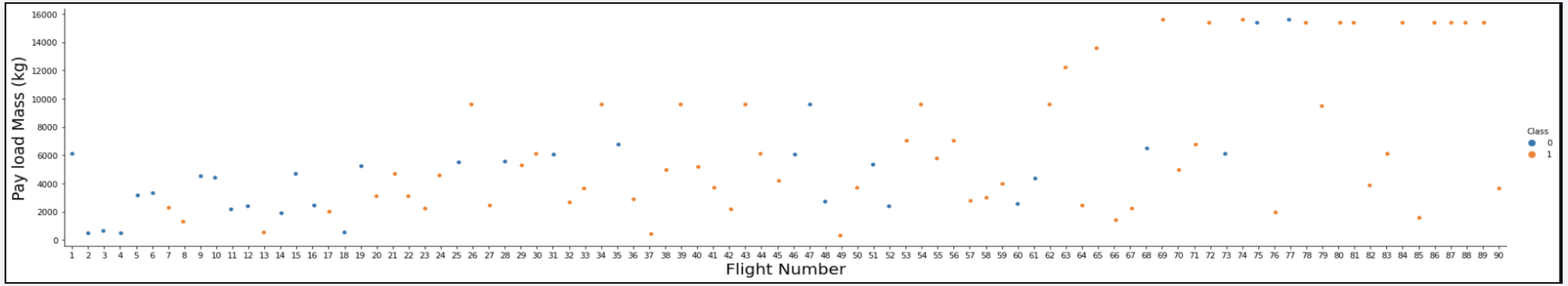
Data Wrangling

- Transform raw data into a clean dataset to provide meaningful data.
- Exploratory Data Analysis (EDA) was performed with an API on the dataset for sampling the data while dealing with Nulls.
- Using API to target the endpoint to gather specific data for each ID number in the case of;
  - Booster
  - Launchpad
  - Payload and
  - Core (Landing Outcome)

- Source code: https://github.com/gotnerd/Space-X-Falcon-9-First-Stage-Landing-Prediction-Lab-2-Data-wrangling/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Exploring the data with the use of scatterplots & barplots for visualization of relationship between pairing of listed features:
    - Flight number VS Payload mass, Flight number VS Launch Site, Payload and Launch Site, Success rate of each Orbit, Flight number & Orbit type, Payload & Orbit type,



- Source link: https://github.com/gotnerd/EDA-with-Visualization-Lab/blob/main/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

The following queries were performed & displayed:

- Names of the unique luanch sites in the space mission.
- Display 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 verstion F9 v1.1
- Date when the first successful landing outcome in ground pad was achieved.
- Names of the boosters which have success in drone ship & have payload mass greater than 4000 but less than 6000.
- Total number of successful & failure mission outcomes.
- Names of the booster versions whih have carried the max payload mass.
- Records which will display the month names, failure lnading outcomes in drone ship, booster versions, launch site for the months in year 2015.
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 & 2017-03-20, in descending order.
- Source link: https://github.com/gotnerd/EDA-with-SQL-/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

The use of a Folium mapping, markers, circles, lines and marker clusters provided the ability to mark the success/failed launches for each site on the map.

- The previous was used to calculate the distance between a launch site to its proximites in order to find some geographical patterns about launch sites.
  - Markers indicate points of launch sites;
  - Circles indicate highlighted areas around specific coordinates such as NASA Johnson Space Center (JSC) & Vandenberg AFB SLC 4E along with NASA Kennedy Space Center (KSC)
  - Marker Clusters indicates groups of events in each coordinate, like launches in a launch site and;
  - Finally, the use of Lines helped indicate the distances between two coordinates on the map.

- Source link: https://github.com/gotnerd/Launch-Sites-Locations-Analysis-with-Folium/blob/main/Launch%20Sites%20Locations%20Analysis%20with%20Folium%20SpaceX%20Project.ipynb

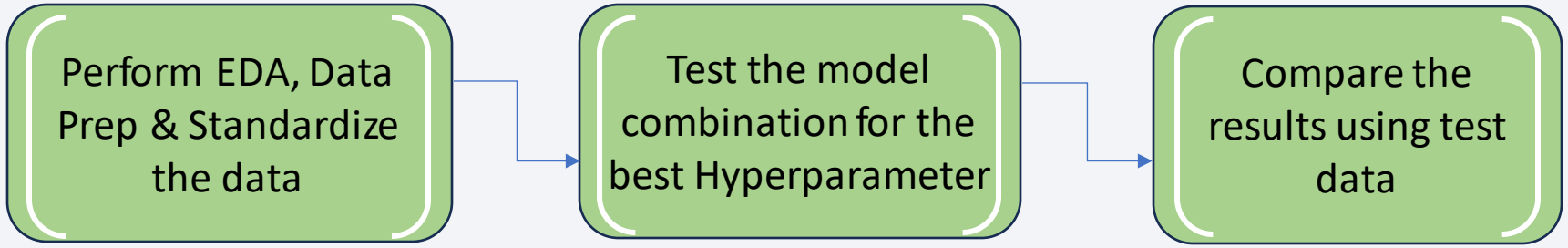
Build a Dashboard with Plotly Dash

- The pie chart and a scatter point chart below are used for visualization data
  - Percentage of launches by site
  - Payload range
- Build a Dashboard Application using Plotly Dash application for users to perform interactive visual analytics on SpaceX launch data in real-time.
- This dashboard application contains input components such as a dropdown list and a range slider to interact

- Source Link: https://github.com/gotnerd/SpaceX-Build-an-Interactive-Dashboard-with-Ploty-Dash/blob/main/dashboard.md

Predictive Analysis (Classification)

- Build a machine learning pipeline to predict if the first stage will land given the data.
- Perform Exploratory Data Analysis and determine Training Labels
  - Create a column for the class
  - Standardize the data
  - Split into training data and test data
- Find best Hyperparameter for SVM, Classification Trees and Logistic Regression
- Find the method performs best using test data



- Source Link: https://github.com/gotnerd/SpaceX-Machine-Learning-Prediction.ipynb/blob/main/SpaceX%20Machine%20Learning%20Prediction.ipynb

Results

Exploratory data analysis results

- Logistic Regression, SVM, Decision Tree & KNN models all achieved an accuracy of 83.33% ON THE TEST DATA.
- All models faced a challenge with false positives in their confusion matrices.
- Further analysis & fine tuning are required to mitigate false positives & enhance overall model performance.
- SpaceX uses 4 different launch sites
- First launches were done by SpaceX and NASA
- Average payload of F9 v1.1 booster is 2,928kg
- Interpretative Results found:
  - True Positives (TP): The model correctly predicted 12 successful first stage landing outcomes.
  - False Positives (FP): The model incorrectly predicted 3 successful landing outcomes, they were not.
  - True Positives (TP): The model correctly predicted unsuccessful landings 3 times.
  - False Positives (FP): The model incorrectly predicted 0 unsuccessful landings outcomes where there was success.
- Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average
- Two booster versions failed at landing in drone ship in 2015: F9 v1.1 B1012 & F9 v1.1 B1015

Results

Findings

- The models excel in correctly identifying successful landings (high True Positive rate).
- The major issue lies in the occurrence of falsse positives, where it predicts a successful landing when it's not.
- Further tuning is needed to reduce this issue & improve overall performance.
- The model have a total accuracy of 83.33% on the test data.

Key Takeaways

- We employed four machine learning models: Logistic Regression, SVM, Decision Trees & KNN.
- All models demonstrated an identical accuracy of 83.33%on the test data.
- The analysis highlighted a common issue in the form of false positives present in all models.

Implications

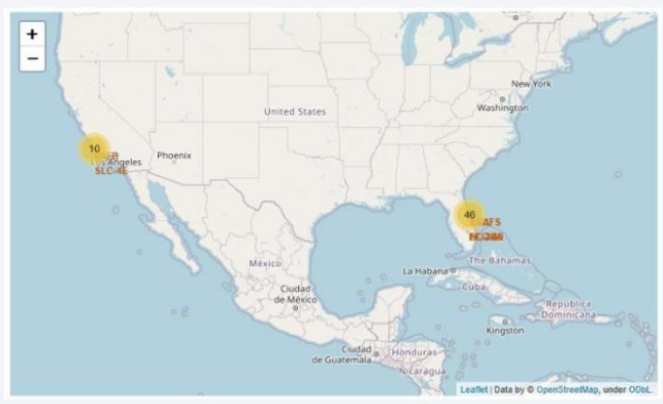
- Predictive models play a vital role in estimating launch costs.
- Addressing the hcallenge of the false positives is crucial for enhancing cost prediction accuracy.

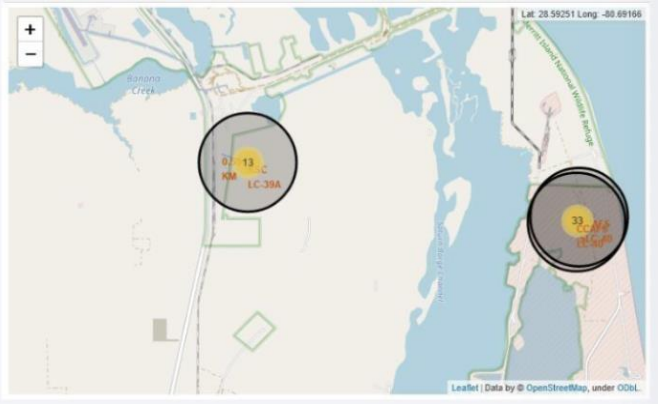
Next Steps

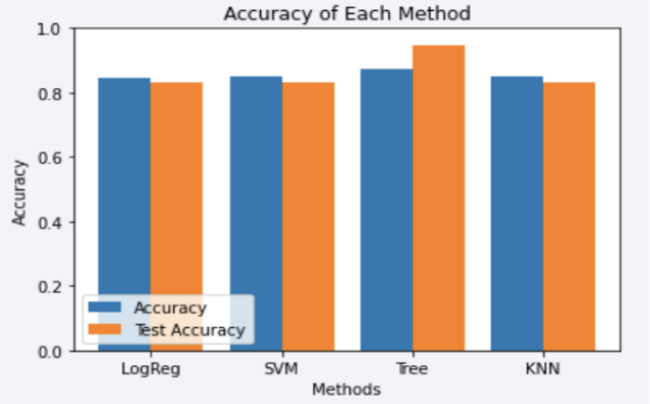
- Further analysis along with fine tuning are essential to reduce the occurrence of false positives.
- Exploring advanced modeling techniques could lead to more reliable predictions.

Results

- The use of interactive analytics was possible to identify which launch sites are in coastal zones & possess sound logistic infrastructure.
- Most launches occur on the East coast, with few on the West coast of the United States. (As shown in Fig's. 1 & 2)



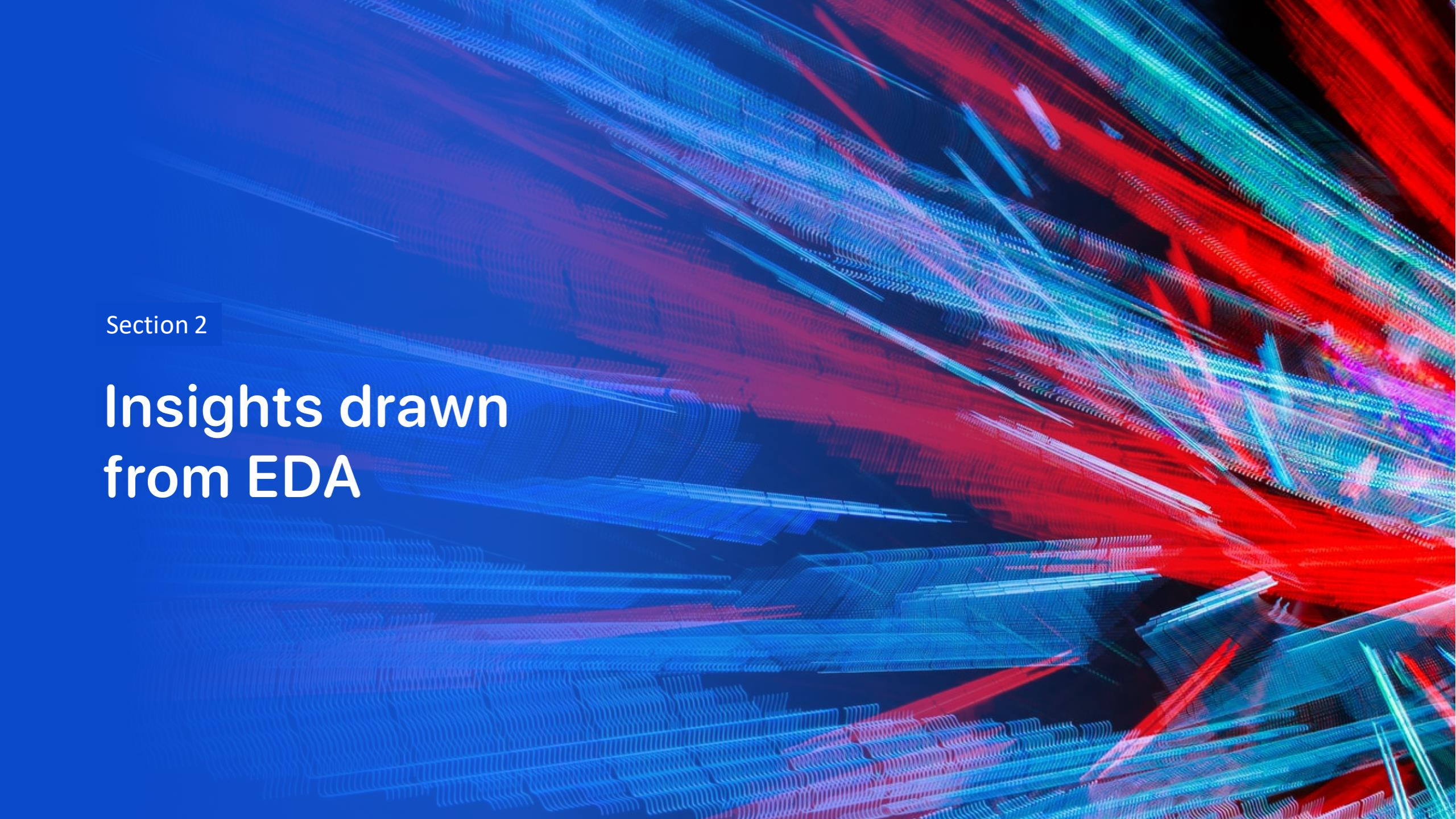




- Predictive Analysis showed that the Decision Tree Classifier is the best model to predict successful landings, having accuracy of 83.33% along with higher accuracy for test data of 94%. (As shown in Fig.3)

Section 2

Insights drawn from EDA



Flight Number vs. Launch Site

- Show a scatter plot of Flight Number vs. Launch Site

- Show the screenshot of the scatter plot with explanations

Payload vs. Launch Site

- Show a scatter plot of Payload vs. Launch Site

- Show the screenshot of the scatter plot with explanations

Success Rate vs. Orbit Type

- Show a bar chart for the success rate of each orbit type

- Show the screenshot of the scatter plot with explanations

Flight Number vs. Orbit Type

- Show a scatter point of Flight number vs. Orbit type

- Show the screenshot of the scatter plot with explanations

Payload vs. Orbit Type

- Show a scatter point of payload vs. orbit type

- Show the screenshot of the scatter plot with explanations

Launch Success Yearly Trend

- Show a line chart of yearly average success rate

- Show the screenshot of the scatter plot with explanations

All Launch Site Names

- Find the names of the unique launch sites
- Present your query result with a short explanation here

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- Present your query result with a short explanation here

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- Present your query result with a short explanation here

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- Present your query result with a short explanation here

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- Present your query result with a short explanation here

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

- Present your query result with a short explanation here

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- Present your query result with a short explanation here

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Present your query result with a short explanation here

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

- Present your query result with a short explanation here

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- Rank 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

- Present your query result with a short explanation here


Section 3
Launch Sites
Proximities Analysis

<Folium Map Screenshot 1>

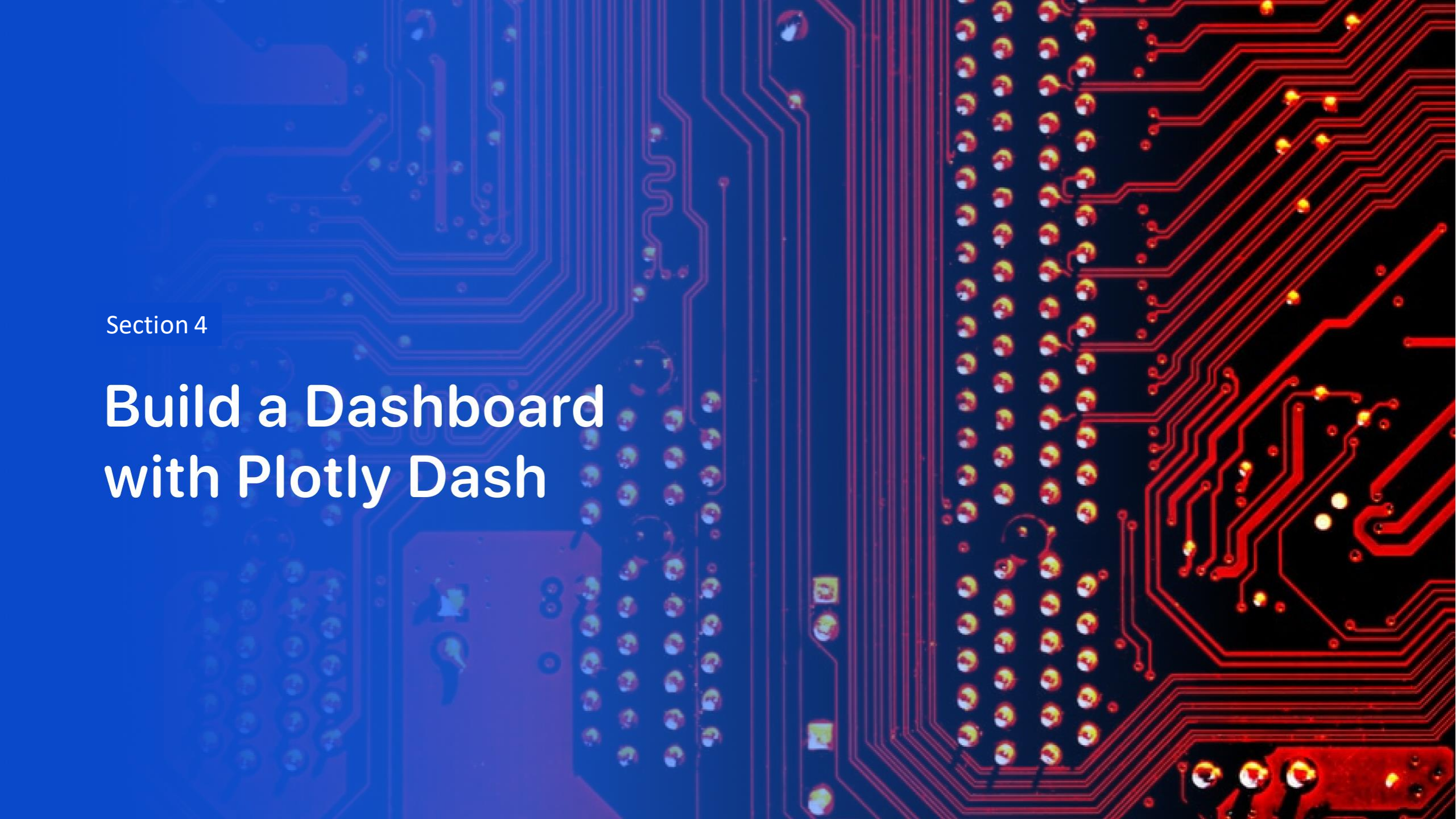
- Replace <Folium map screenshot 1> title with an appropriate title
- Explore the generated folium map and make a proper screenshot to include all launch sites' location markers on a global map
- Explain the important elements and findings on the screenshot

<Folium Map Screenshot 2>

- Replace <Folium map screenshot 2> title with an appropriate title
- Explore the folium map and make a proper screenshot to show the color-labeled launch outcomes on the map
- Explain the important elements and findings on the screenshot

<Folium Map Screenshot 3>

- Replace <Folium map screenshot 3> title with an appropriate title
- Explore the generated folium map and show the screenshot of a selected launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed
- Explain the important elements and findings on the screenshot


Section 4
Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>

- Replace <Dashboard screenshot 1> title with an appropriate title
- Show the screenshot of launch success count for all sites, in a piechart
- Explain the important elements and findings on the screenshot

<Dashboard Screenshot 2>

- Replace <Dashboard screenshot 2> title with an appropriate title
- Show the screenshot of the piechart for the launch site with highest launch success ratio
- Explain the important elements and findings on the screenshot

<Dashboard Screenshot 3>

- Replace <Dashboard screenshot 3> title with an appropriate title
- Show screenshots of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider
- Explain the important elements and findings on the screenshot, such as which payload range or booster version have the largest success rate, etc.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Visualize the built model accuracy for all built classification models, in a bar chart

- Find which model has the highest classification accuracy

Confusion Matrix

- Show the confusion matrix of the best performing model with an explanation

Conclusions

- Point 1
- Point 2
- Point 3
- Point 4
- …

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!