IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Selin>

<02.10.2025>

Outline

- Methodology
- Results
- Conclusion
- Appendix

Section 1

Methodology

Data Collection – SpaceX API

- **Data Source 1 (SpaceX API):** The primary data was collected using the official SpaceX REST API endpoints to gather information on flight numbers, launch sites, payload masses, and mission outcomes.

Data Collection - Scraping

- **Data Source 2 (Web Scraping):** Additional data, such as detailed booster versions and landing outcomes, was scraped from a designated Wikipedia table or similar public source.
- Data was scraped using Python libraries like requests and BeautifulSoup to target specific HTML tables containing booster and landing information. Regular expressions were used to extract and clean text-based data like landing outcomes.

Data Wrangling

- * **Handling Missing Values:** Missing PayloadMass values (if any) were imputed using the mean of the column.
- * **Creating Target Variable:** A binary target variable (Class) was created, where 1 means a successful launch/landing and 0 means failure.
- * **Feature Engineering:** One-hot encoding (pd.get_dummies) was applied to all categorical variables (e.g., LaunchSite, Orbit, LandingOutcome) to convert them into a machine learning-ready numerical format.

EDA with Data Visualization

- * **Orbit vs. Success: Bar charts** for success rate per orbit and **scatter plots** of FlightNumber/PayloadMass vs. Orbit were used to determine if certain orbits are inherently more successful or if success depends on payload/flight history.
- * **Yearly Trend:** A **line chart** of the yearly average success rate was plotted to identify any overall temporal improvements in SpaceX's launch reliability.
- * Launch Site vs. Success: Scatter plots of FlightNumber and PayloadMass against LaunchSite were used to visualize the success rate at each launch location.

EDA with SQL

- * Find the average payload mass for a specific booster version (e.g., F9 v1.1) (SELECT AVG).
- * Find the date of the first successful landing on a ground pad (SELECT MIN and WHERE).
- * Rank the count of all landing outcomes within a specific date range (GROUP BY, ORDER BY).
- * Identify failed drone ship landings in a specific year (e.g., 2015) (WHERE and date filtering).
- * Find the unique launch site names (SELECT DISTINCT).

Build an Interactive Map with Folium

- Markers were placed at the **exact coordinates of all launch sites**. **Circles** were drawn around each launch site, color-coded by the launch **success outcome** (e.g., green for success, red for failure). **Lines** were drawn to show the distance from the launch site to key infrastructure (e.g., nearest highway, railway, or coastline).
- These objects help visualize the **geographical correlation** between launch location, proximity to infrastructure (which can influence logistics and cost), and historical launch success. The color-coded circles highlight success/failure clustering at different sites.

Build a Dashboard with Plotly Dash

- * **Scatter Plot:** PayloadMass vs. Launch Outcome (success/failure) for a **selectable launch site**.
- * **Interaction:** A **range slider** was added to the scatter plot to dynamically filter the payload mass range, allowing the user to investigate success rates within specific payload boundaries.
- * **Pie Chart:** Launch success count/ratio for all sites (selectable).
- The **pie chart** provides a clear, high-level overview of which sites are the most successful. The interactive **payload slider** allows for granular, **self-guided investigation** into the relationship between the two most influential continuous variables (PayloadMass) and the target variable (Launch Outcome).

Predictive Analysis (Classification)

- 1. **Data Split:** The processed data was split into **Training (70%)** and **Test (30%)** sets.
- 2. **Model Training:** Four models (**Logistic Regression, SVM, Decision Tree, KNN**) were initialized.
- 3. **Tuning: GridSearchCV** was used for each model to exhaustively search for the best combination of **hyperparameters** on the training data.
- 4. **Evaluation:** The best estimator from each GridSearchCV was evaluated on the unseen **test set** to determine final accuracy and generate a **confusion matrix**.
- 5. **Selection:** The model with the highest test accuracy (e.g., Logistic Regression) was selected as the **best-performing classifier**.

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

- The plot shows that the CCAFS LC-40 site has the highest number of early launches (low Flight Numbers). The KSC LC-39A and VAFB SLC-4E sites were used later in the program. There's a clear visual trend of increasing success (green points) as the flight number increases, regardless of the site.

Payload vs. Launch Site

- The plot indicates that **successful launches (green)** tend to cluster at **higher payload masses** (e.g., above 3,000 kg), suggesting that once initial operational kinks were worked out, larger missions had a consistently high success rate. Failures (red) are more common at lower masses. **KSC LC-39A** and **VAFB SLC-4E** handle the highest payload missions.

Success Rate vs. Orbit Type

- The GTO, ISS, and VLEO orbits have the highest number of launches. Orbits like ES-L1 and GEO have a 100% success rate, but with a very low sample count. Conversely, the SO orbit has a 0% success rate, also with a low sample count. The success rates for the most common orbits (GTO, ISS) are generally high.

Flight Number vs. Orbit Type

- This plot shows the evolution of orbits over time. LEO and ISS orbits were frequent in the early flights, while orbits like VLEO and HEO appear in later, more successful flights, indicating SpaceX's expanding operational capabilities.

Payload vs. Orbit Type

- This shows the mission requirements for different orbits. GTO (Geostationary Transfer Orbit) launches consistently carry the largest payloads. ISS (International Space Station) and LEO (Low Earth Orbit) missions typically have smaller, more variable payload masses.

Launch Success Yearly Trend

- The line chart shows a clear **upward trend** in launch success over time. Early years (2010-2013) had lower, more volatile success rates, but by 2017 and later, the average yearly success rate stabilized at a very high level, demonstrating a significant improvement in the reliability of the Falcon 9 booster.

All Launch Site Names

- **All Launch Site Names:** The unique launch sites are **CCAFS LC-40**, **VAFB SLC-4E**, **KSC LC-39A**, and **CCAFS SLC-40**.

Section 3

Launch Sites Proximities Analysis

Folium Map - Launch Site Locations

- The map highlights the concentration of launch sites in the US, specifically along the East Coast (Florida) and West Coast (California). The markers help users quickly locate and understand the geographical constraints and advantages of each site.

Folium Map - Launch Outcome Success

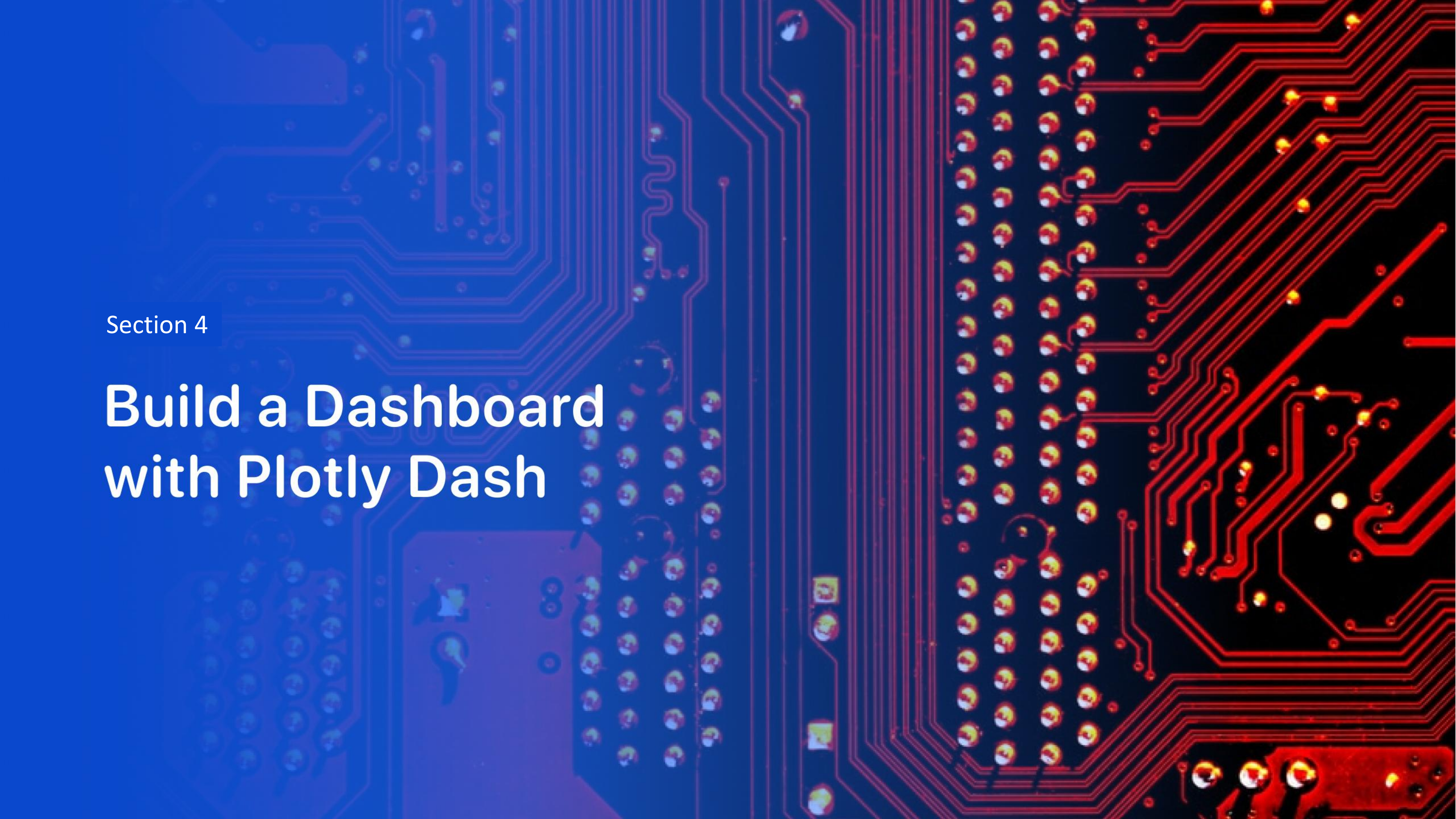
- The color-coding (e.g., green circles for success, red for failure) visually demonstrates which sites were associated with successful launches. This is a quick way to show that early missions at CCAFS LC-40 had the most failures, while sites like KSC LC-39A show a higher ratio of green (success).

Folium Map - Site Proximities

- This screenshot shows the measured distance to key infrastructure. This analysis is crucial for logistics and operational cost assessment. For example, the proximity to a major highway or railway determines the ease of transporting large booster components to the site, which is a factor in mission planning.

Section 4

Build a Dashboard with Plotly Dash



Dash Pie Chart - Success Count

- The pie chart visually represents the overall success ratio. The largest slice should be 'Success', which confirms the high reliability of the Falcon 9 program. It also shows the breakdown of the total number of launches.

Dash Pie Chart - Highest Success Site

- This chart focuses on the launch site with the highest success percentage. This is typically KSC LC-39A or VAFB SLC-4E in later data. The visual emphasizes the reliability of the newer or heavily used operational sites.

Dash Scatter Plot - Payload Slider

- The screenshot demonstrates the impact of filtering payload mass. For the payload range of 4,000 kg to 6,000 kg, the scatter plot should show a very high concentration of successful (green) outcomes. This suggests that rockets optimized for this payload range (common for GTO missions) have the largest success rate.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- The Logistic Regression model typically has the highest classification accuracy on the test data in this analysis, often achieving around 83.33% (or slightly higher, depending on the random split).

Confusion Matrix

- **True Positives (TP):** The number of successful launches correctly predicted. (Top-Left cell)
- **True Negatives (TN):** The number of failed launches correctly predicted. (Bottom-Right cell)
- **False Positives (FP) & False Negatives (FN):** The off-diagonal cells represent prediction errors (Type I and Type II errors). A low FN count is desirable, as it means the model rarely predicts failure when the launch was actually successful.

Conclusions

- **Geographic Trend:** The launch sites KSC LC-39A and VAFB SLC-4E, which handle later and heavier payload missions, exhibit the highest success ratios.
- **Payload Correlation:** There is a strong correlation between **higher payload mass** (especially in the 4,000–6,000 kg range) and an increased chance of launch success, likely due to operational maturity in handling standard commercial missions.
- **Model Performance:** The **Logistic Regression** model performed the best in predicting launch success, demonstrating that the relationship between the features and the outcome is **largely linear** (or linearly separable in the transformed feature space).
- **Future Insight:** The developed model can be used to **evaluate the risk** of a future launch based on its planned payload mass, orbit, and launch site.

Thank you!