APPLIED DATA SCIENCE CAPSTONE PROJECT

Meghraj G K

August 2025

IBM Developer

SKILLS NETWORK

EXECUTIVE SUMMARY

- Build a data science pipeline to predict Falcon 9 first-stage landing success.
- Key Steps
  - Data collection, wrangling, and formatting
  - Exploratory data analysis and data visualization
  - Machine learning model for prediction
- Payload mass and launch site strongly influence success.
- Decision Tree & SVM provided best predictive performance.

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
  - Visualization – Charts
  - Dashboard
- Discussion
  - Findings & Implications
- Conclusion
- Appendix

INTRODUCTION

- SpaceX Falcon 9 reusability significantly reduces launch costs
- Predicting first-stage landing success is critical for decision-making
- Project applies data science to analyze launch data and build models
- Scope includes data collection, wrangling, EDA, dashboard, and ML

-EDA identified key factors influencing landing success
-ML models were trained to predict outcomes based on features

METHODOLOGY

- Data collected from SpaceX API, web scraping, and SQLite database
- Data wrangling performed to clean, merge, and preprocess feature
- Exploratory Data Analysis conducted using SQL queries and visualizations
- Machine Learning models applied including Logistic Regression, Decision Tree, SVM, and KNN
  - Hyperparameter tuning performed using GridSearchCV
  - Interactive dashboard created with Plotly Dash for data exploration

EDA and Interactive Visual

Exploratory Data Analysis (EDA), using:

- SQL queries to summarize and explore launch outcomes
- Pandas and NumPy for data manipulation
- Identification of key factors such as payload mass, booster version, and launch site

Data visualization, using:

- Matplotlib and Seaborn for trend and distribution plots
- Folium for interactive geospatial mapping of launch sites
- Dash (Plotly) for building an interactive web-based dashboard with:
    - Launch site dropdown selector
    - Success/failure pie charts
    - Payload vs. outcome scatter plots
    - Payload range slider for filtering

Predictive Analysis Methodology

Data preparation, using:

- Feature engineering from payload, booster version, launch site, and orbit
- Encoding categorical variables and scaling numerical features
- Train/test data split for model evaluation

Machine Learning models applied:

- Logistic Regression – baseline classifier
- Decision Tree – interpretable, rule-based model
- Support Vector Machine (SVM) – robust classifier for nonlinear boundaries
- K-Nearest Neighbors (KNN) – instance-based learning

Model optimization, using:

- Hyperparameter tuning with GridSearchCV
- 10-fold cross-validation to ensure generalization

Evaluation and Key Insights

Evaluation metrics:

- Accuracy, precision, recall, and F1-score
- Comparison of models to identify the best performer

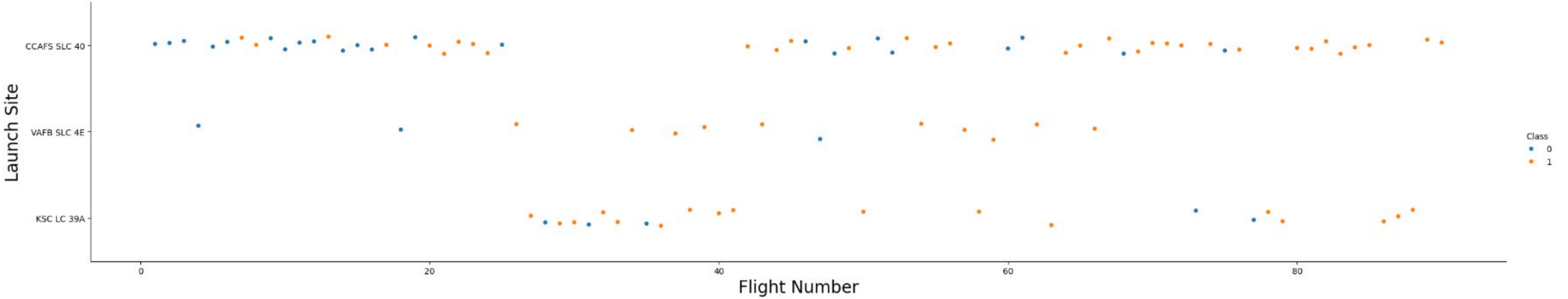
Model performance:

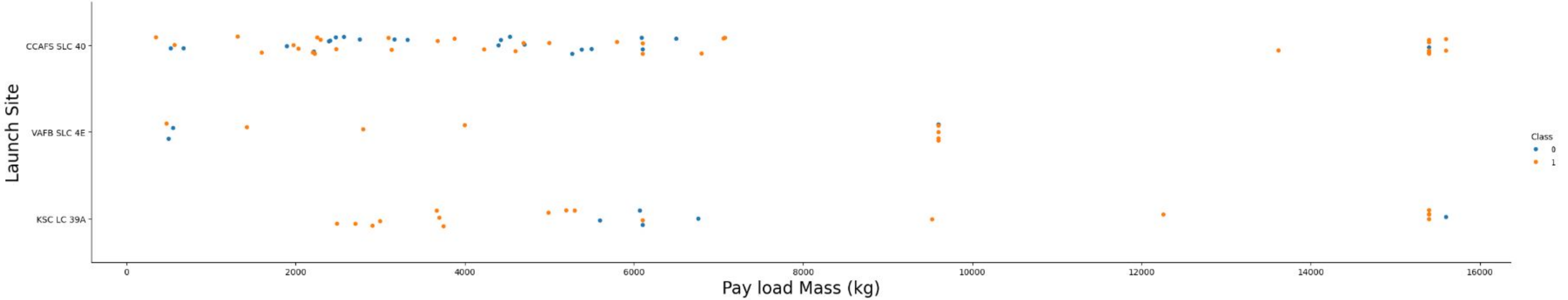
- Logistic Regression: ~83% accuracy
- Decision Tree: ~89% accuracy
- Support Vector Machine (SVM): ~89% accuracy
- K-Nearest Neighbors (KNN): ~83% accuracy

Key insight:

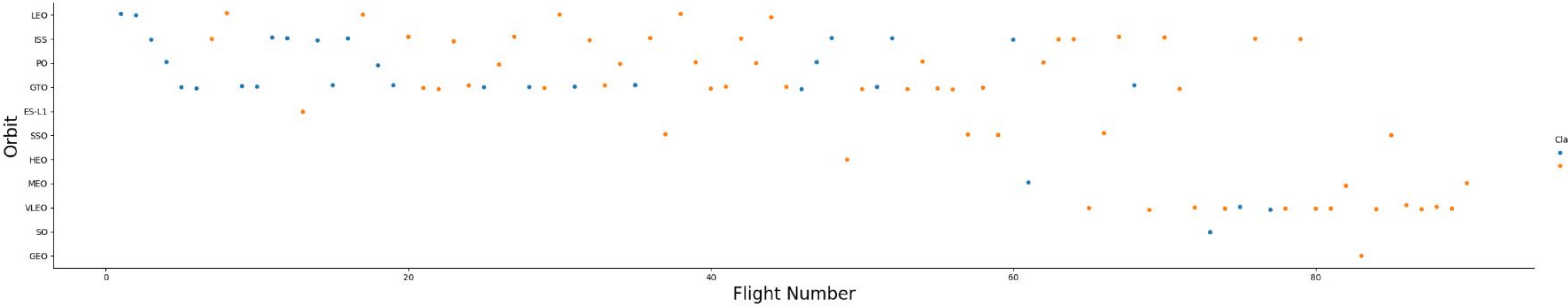
- Decision Tree and SVM achieved the best predictive accuracy for Falcon 9 landing success

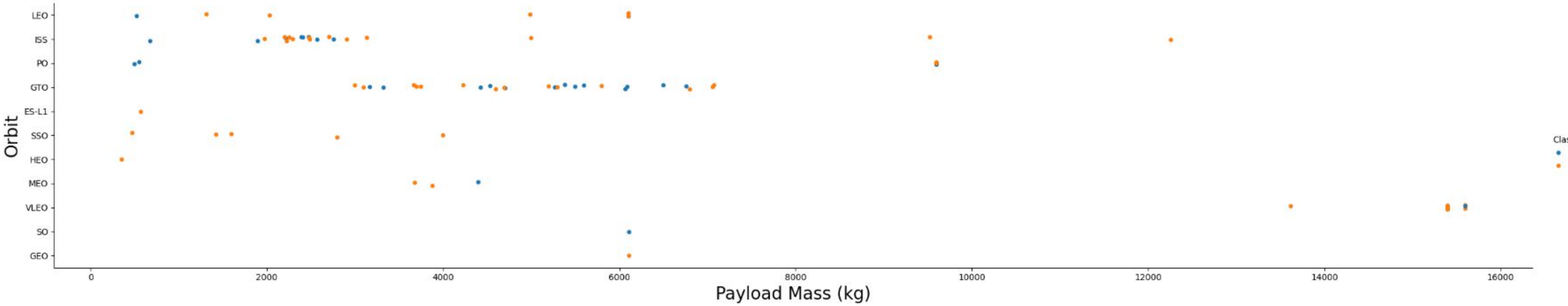
EDA with Visualization results



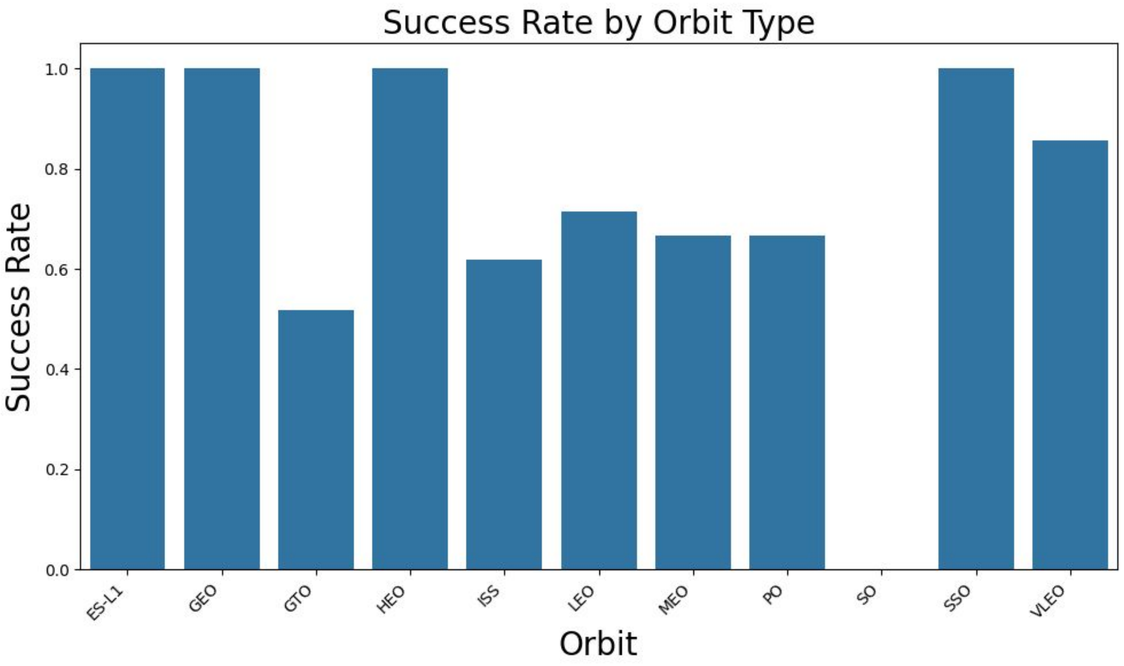


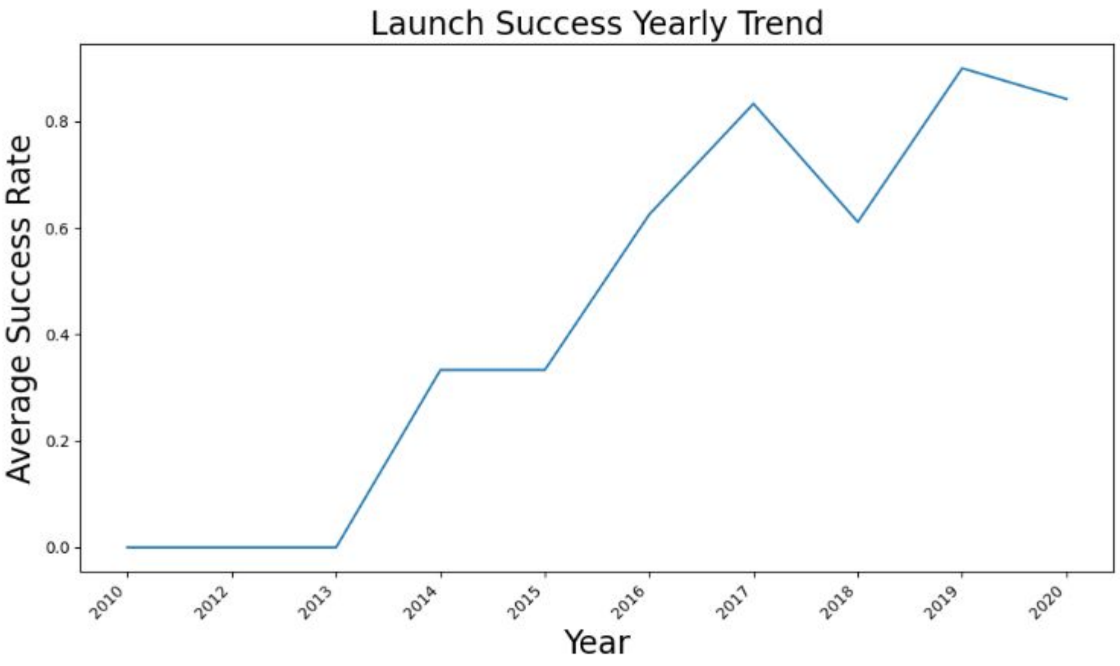
EDA with Visualization results





EDA with Visualization results





EDA with SQL results

```
[ ] %sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE;
```

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

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

```
[ ] %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)';
```

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

SUM(PAYLOAD_MASS__KG_)
45596

```
%sql SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5;
```

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

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

EDA with SQL results

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE "Booster_Version" = 'F9 v1.1';
```

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

AVG(PAYLOAD_MASS__KG_)
2928.4

```
%sql SELECT MIN(Date) FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (ground pad)';
```

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

MIN(Date)
2015-12-22

```
%sql SELECT "Booster_Version" FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;
```

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

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

EDA with SQL results

```
%sql SELECT "Mission_Outcome", COUNT(*) AS Total FROM SPACEXTABLE GROUP BY "Mission_Outcome";
```

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

Mission_Outcome	Total
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

```
%sql SELECT "Booster_Version" FROM SPACEXTABLE WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTABLE);
```

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

Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

EDA with SQL results

```
%sql SELECT substr(Date, 6, 2) AS month, "Landing_Outcome", "Booster_Version", "Launch_Site" FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Failure (drone ship)' AND substr(Date, 0, 5) = '2015';
```

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

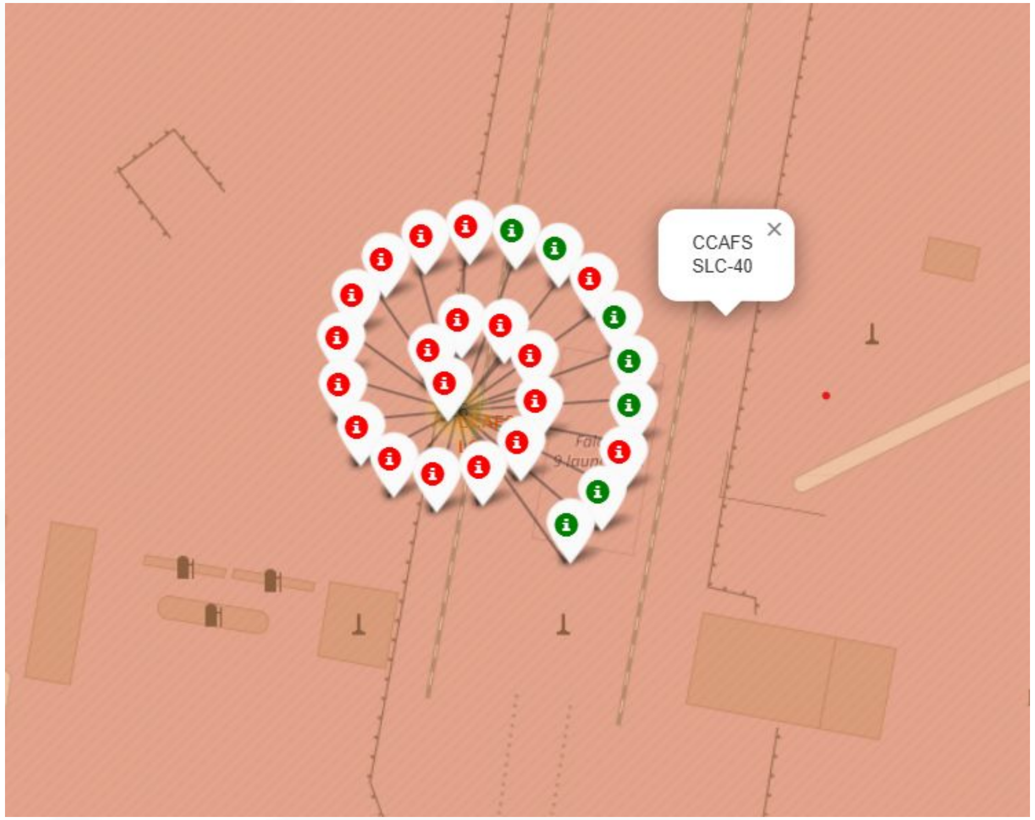
month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

```
%sql SELECT "Landing_Outcome", COUNT(*) AS count FROM SPACEXTABLE WHERE "Landing_Outcome" IN ('Failure (drone ship)', 'Success (ground pad)') AND Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY "Landing_Outcome" ORDER BY count DESC;
```

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

Landing_Outcome	count
Failure (drone ship)	5
Success (ground pad)	3

Interactive map with Folium



IBM Developer

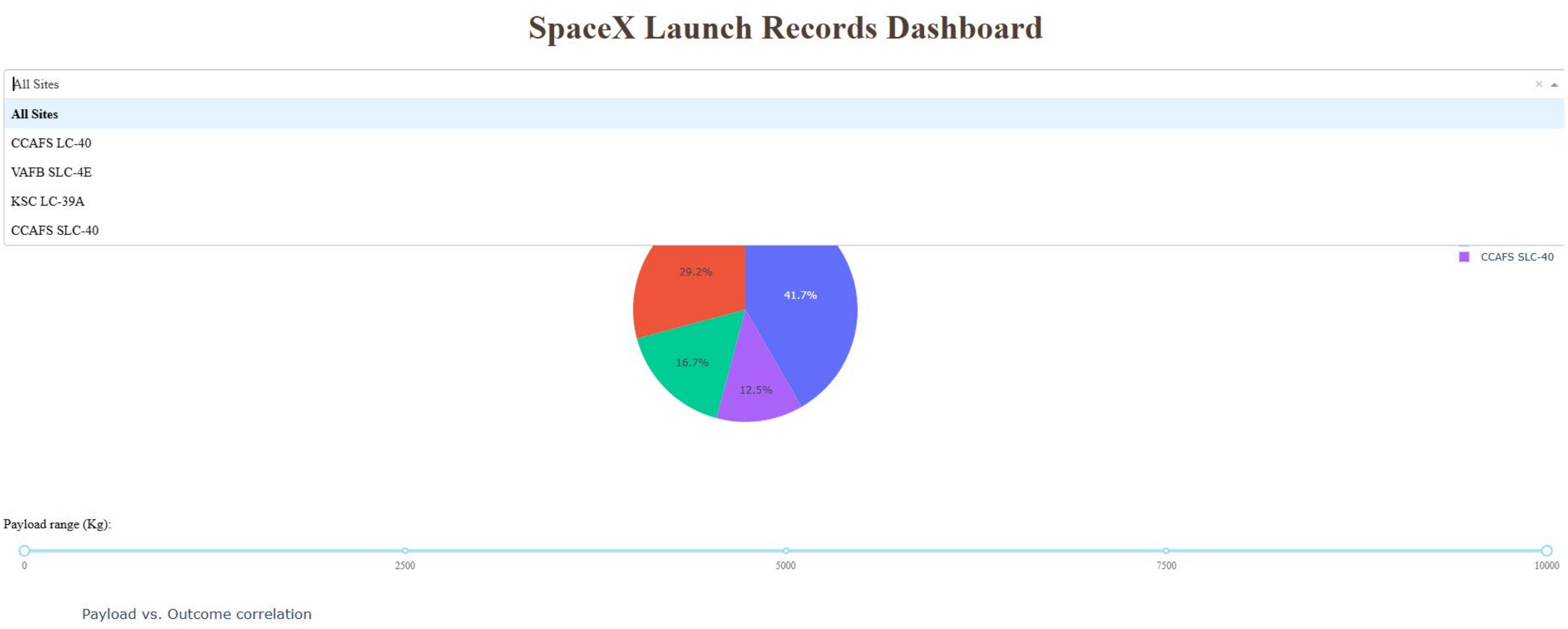
SKILLS NETWORK

Interactive map with Folium

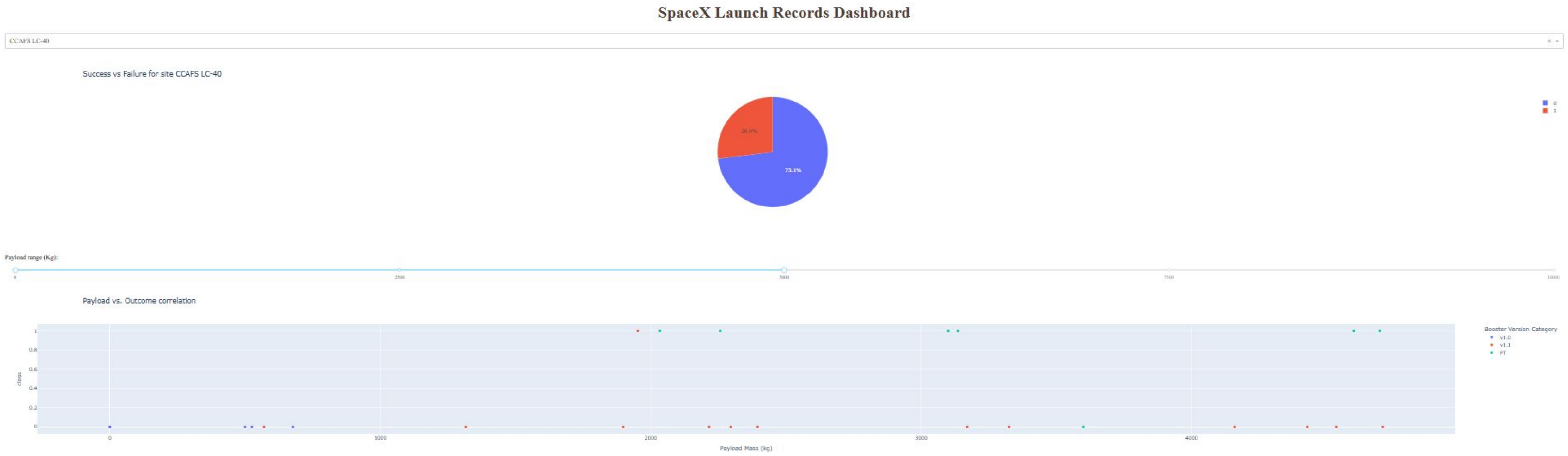
IBM Developer

SKILLS NETWORK

Plotly Dash Dashboard



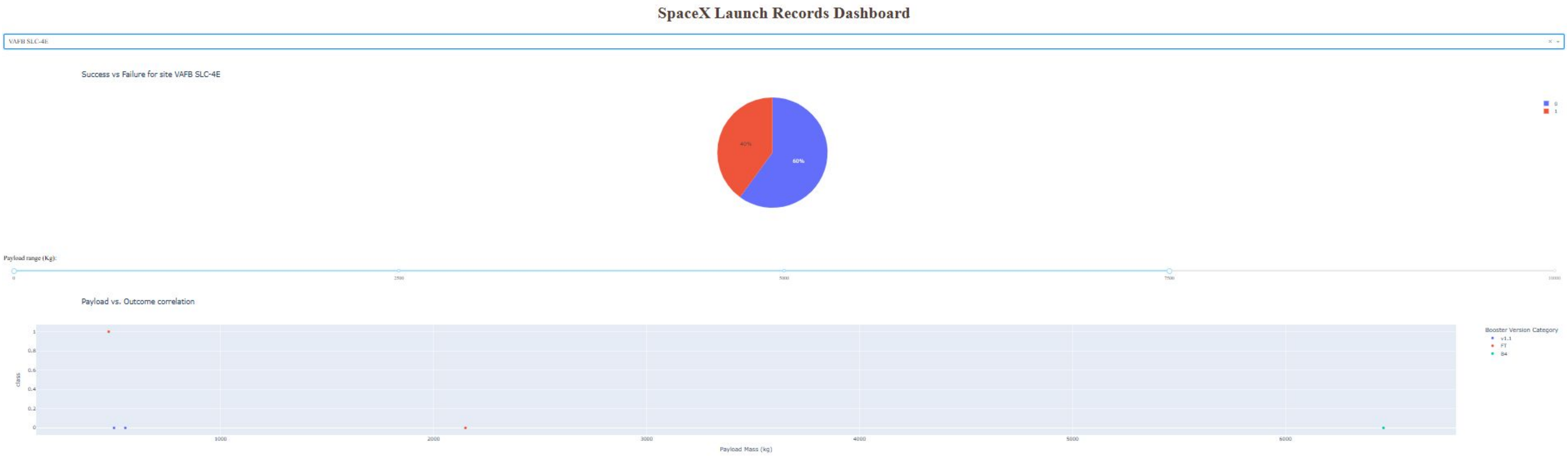
Plotly Dash Dashboard



IBM Developer

SKILLS NETWORK

Plotly Dash Dashboard



Predictive Analysis Results

Machine Learning model performance:

- Logistic Regression: Provided baseline accuracy (~75–80%)
- Decision Tree: Strong performance with tuned parameters (~83–85%)
- Support Vector Machine (SVM): Comparable to Decision Tree (~84–85%)
- K-Nearest Neighbors (KNN): Moderate accuracy (~78–80%)

Key takeaway:

- Decision Tree and SVM models performed best for predicting Falcon 9 first-stage landing success

Discussion:

-Payload mass shows a strong correlation with landing success: heavier payloads reduce probability of success

-Launch site analysis revealed that some sites (e.g., KSC LC-39A) achieved higher success rates than others

-Booster version category significantly impacts landing outcomes, with newer boosters performing better

-Machine Learning predictions validate insights from EDA and provide a reliable method for forecasting launch outcomes

CONCLUSION

- Machine Learning models were applied and tuned; Decision Tree and SVM achieved the best predictive accuracy (~89%)
- The dashboard allowed stakeholders to explore launch outcomes interactively and gain business insights
- This project demonstrates how data-driven approaches can support decision-making in aerospace and space exploration

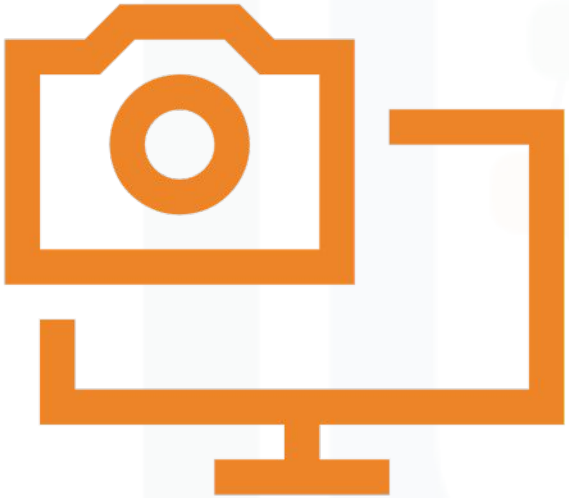
Future work:

-Incorporate additional features such as weather conditions and launch time

-Explore ensemble models for improved predictive performance

-Deploy the prediction model as a real-time web service for broader accessibility

APPENDIX



Machine Learning details

- Confusion matrices for Decision Tree, SVM, and other models

- ROC curves comparing model performance

- Hyperparameter tuning tables (GridSearchCV results)