SPACEX LAUNCH DATA ANALYSIS AND LANDING PREDICTION

Data Science Capstone Project

Gautam Murmu

Course – Applied Data Science Capstone

CORE VALUES

This project reflects core data science values including curiosity to explore real-world data, analytical thinking to uncover meaningful patterns, and effective communication to present insights clearly. Throughout this capstone, I applied data collection, data wrangling, visualization, and machine learning techniques to solve a practical problem and present the findings in a structured and impactful manner.

EXECUTIVE SUMMARY

This project analyzes SpaceX launch data to understand the factors that influence successful rocket landings and to identify patterns across launch sites. The analysis began with data collection from APIs and datasets, followed by data wrangling and exploratory data analysis using visualizations and SQL queries. Interactive visual analytics were performed using Folium maps and a Plotly Dash dashboard to better understand geographical and operational patterns.

Finally, multiple machine learning classification models including Logistic Regression, Support Vector Machine, Decision Tree, and K-Nearest Neighbours were trained to predict landing outcomes. The results demonstrate how data science techniques can be applied end-to-end to extract insights and build predictive models from real-world aerospace data.

INTRODUCTION

- SpaceX has significantly reduced the cost of space travel by developing reusable rockets. A key factor in reusability is the successful landing of the first stage booster after launch.
- This project focuses on analyzing historical SpaceX launch data to understand the patterns behind successful and unsuccessful landings. By examining launch site locations, flight data, payload information, and landing outcomes, we aim to discover insights that explain what contributes to a successful landing.
- The project also demonstrates how data science tools such as data wrangling, visualization, SQL analysis, interactive dashboards, and machine learning models can be combined to solve a real-world prediction problem.

DATA COLLECTION & DATA WRANGLING

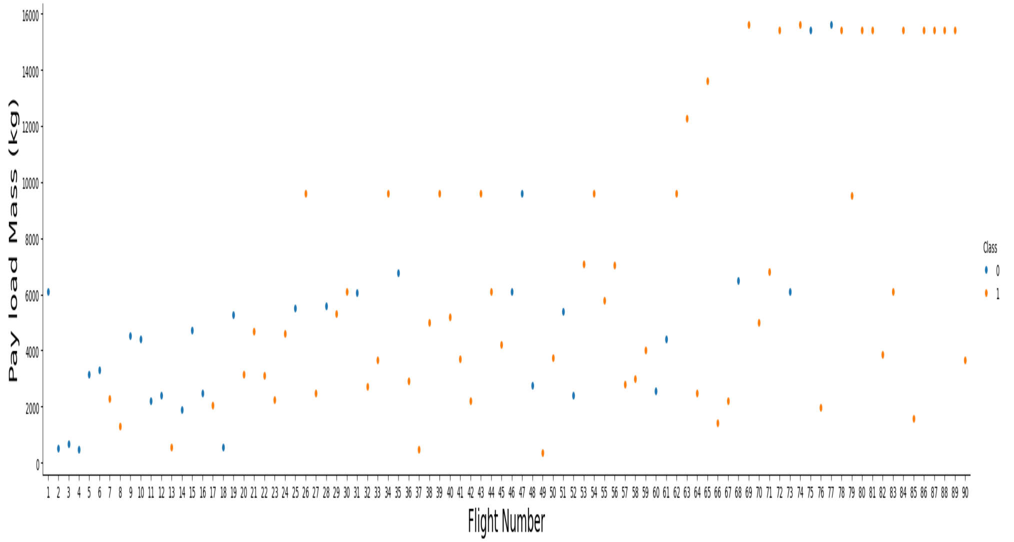
The dataset for this project was collected from SpaceX launch records and publicly available data sources. The raw data contained multiple attributes related to launch site, payload mass, orbit, booster version, and landing outcomes.

- During the data wrangling phase, the dataset was cleaned and prepared for analysis by:
- Handling missing values
- Removing irrelevant columns
- Converting categorical data into usable format
- Creating new features required for analysis
- Formatting the target variable (Class) for prediction

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	La
4	1	2010-06-04T18:45:00.000Z	Falcon 9	5919.165341	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.5
5	2	2012-05-22T07:44:00.000Z	Falcon 9	525.000000	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005	-80.577366	28.5
6	3	2013-03-01T19:10:00.000Z	Falcon 9	677.000000	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007	-80.577366	28.5
7	4	2013-09-29T16:00:00.000Z	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003	-120.610829	34.6
8	5	2013-12-03T22:41:00.000Z	Falcon 9	3170.000000	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004	-80.577366	28.5

EDA WITH VISUALIZATION(MATPLOTLIB/SEABORN)

- Exploratory Data Analysis was performed using visualizations to understand patterns in launch success.
- The analysis revealed relationships between launch site, orbit type, payload mass, and landing outcomes.
- Visual charts helped identify which factors strongly influence the success of first-stage landings.

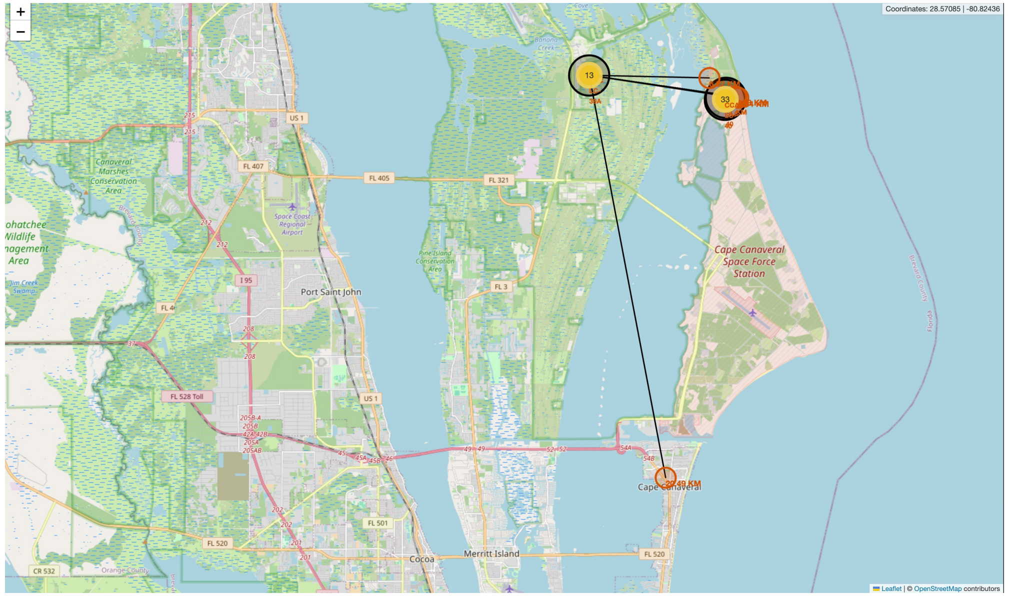


EDA USING SQL QUERIES

- SQL queries were used to extract meaningful insights from the SpaceX launch dataset.
- By querying the database, patterns such as launch site performance, booster reliability, and landing success were identified.
- This step helped in validating patterns observed during visual EDA using structured database analysis.

INTERACTIVE MAP VISUALIZATION USING FOLIUM

- An interactive Folium map was created to visualize SpaceX launch sites geographically.
- Marker clusters were used to represent successful (green) and failed (red) launches.
- Distance lines were drawn from launch sites to nearby coastlines, highways, railways, and cities to analyze geographical advantages.
- This visualization helped understand why launch sites are strategically located near coastlines and away from dense cities.

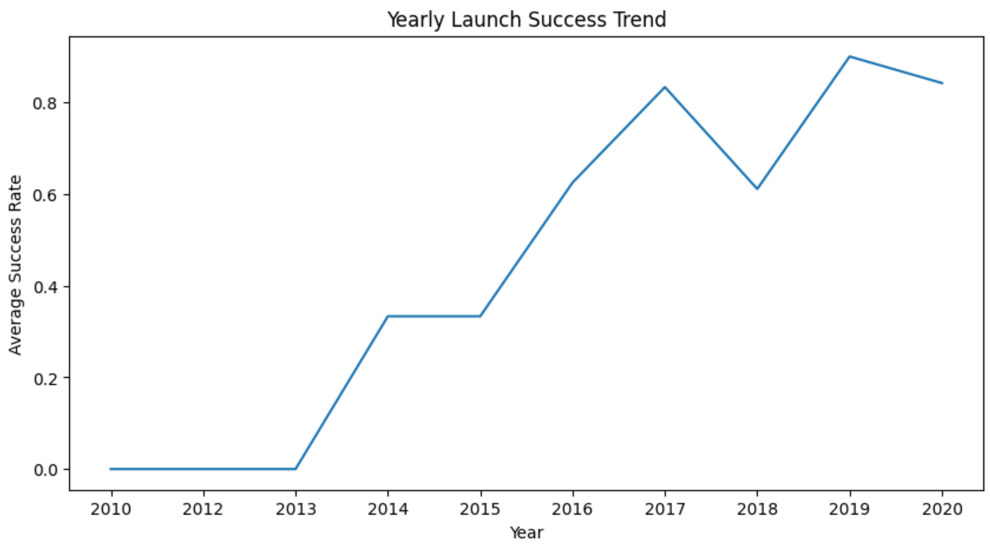


INTERACTIVE DASHBOARD USING PLOTY DASH

This dashboard was built using Plotly Dash to provide interactive visual analysis of SpaceX launch data.

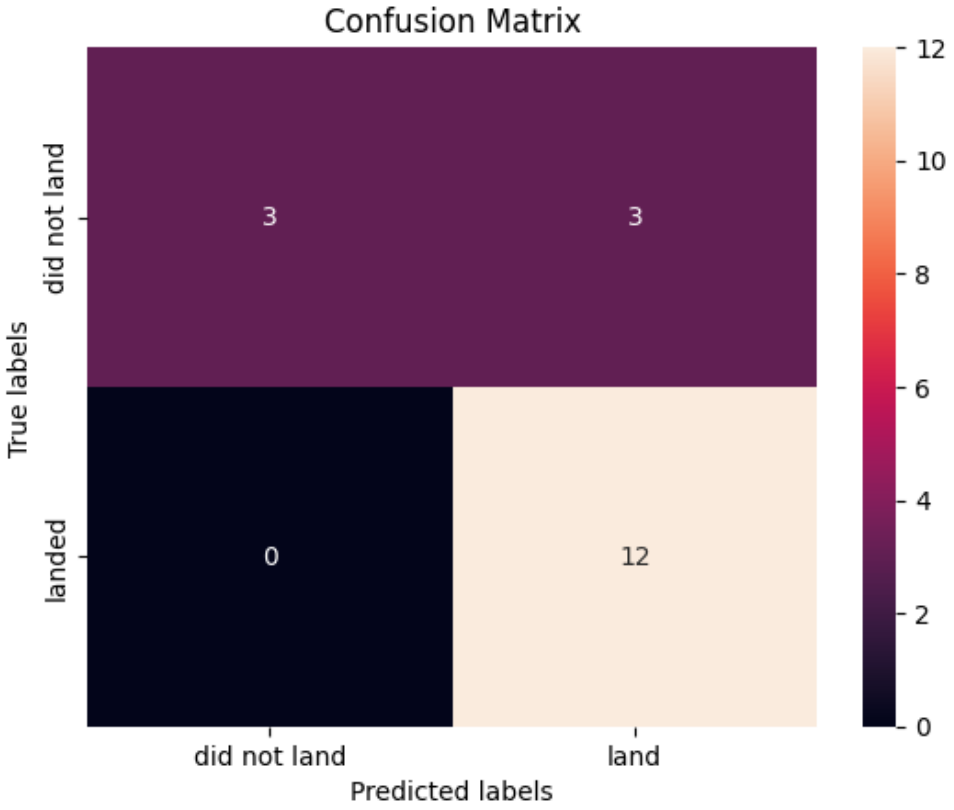
The dashboard allows users to:

- Select different launch sites from a dropdown menu
- View success rate using a pie chart
- Analyze relationship between payload mass and launch outcome using a scatter plot



PREDICTIVE ANALYSIS USING CLASSIFICATION MODEL

- Multiple classification models were built to predict the success of Falcon 9 first stage landing.
- The models used include Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbours (KNN).
- The dataset was split into training and testing sets, and GridSearchCV was used to find the best hyperparameters.
- The performance of each model was evaluated using accuracy score and confusion matrix.



MODEL COMPARISON AND BEST PERFORMING ALGORITHM

- The performance of multiple classification algorithms was compared using the test dataset.
- Logistic Regression, SVM, Decision Tree, and KNN models were evaluated based on accuracy.
- Among all models, the Support Vector Machine (SVM) with sigmoid kernel produced the best validation performance.
- This indicates that SVM was most effective in learning the patterns required to predict the landing outcome.

```
Logistic Regression Test Accuracy: 0.8333333333333334
SVM Test Accuracy: 0.8333333333333334
Decision Tree Test Accuracy: 0.7777777777777778
KNN Test Accuracy: 0.8333333333333334
```

CONCLUSION

- In this project, SpaceX launch data was analysed using data wrangling, visualization, interactive mapping, dashboard development, and machine learning techniques.
- Exploratory data analysis revealed important patterns between payload mass, launch sites, and mission success. Interactive visualizations using Folium and Plotly Dash helped in understanding launch site behaviour and yearly success trends.
- Multiple classification models were built to predict launch success. Logistic Regression, SVM, and KNN achieved the highest accuracy of 83.33%, demonstrating reliable predictive performance.
- This project highlights how data science techniques can be applied to real-world aerospace data to generate insights and build predictive systems.

PROJECT REPOSITORY & REFERENCES

https://github.com/Gautamr06/Applied-Data-Science-Capstone

References:

- IBM Data Science Capstone Labs and datasets
- SpaceX launch data from public APIs
- Folium documentation for interactive maps
- Plotly Dash documentation for dashboard development
- Scikit-learn documentation for machine learning models