IBM DATA SCIENCE CAPSTONE PROJECT

AGENDA

- INTRODUCTION
- OBJECTIVE & IMPORTANCE
- EXECUTIVE SUMMARY
- METHODOLOGY
- RESULTS
- CONCLUSION
- APPENDIX

INTRODUCTION

Space Exploration Technologies Corp. (SpaceX) is an American aerospace manufacturer and space transport services company founded by Elon Musk in 2002. SpaceX is known for its ambitious goals of reducing the cost of space travel and making it possible for humans to live on other planets. The Falcon 9 rocket, a flagship product of SpaceX, is designed for reliable and safe transport of satellites and the Dragon spacecraft into orbit.

OBJECTIVE OF THE PROJECT

The primary objective of this project is to predict the first stage landing outcomes of SpaceX Falcon 9 rockets using a combination of machine learning techniques and data analysis. Successful prediction of these outcomes is crucial for optimizing operational efficiency and enhancing cost-effectiveness in space missions.

IMPORTANCE OF THE PROJECT

This project's significance lies in its comprehensive approach to leveraging data science for aerospace advancement. By meticulously collecting and analysing SpaceX Falcon 9 launch data, including through EDA, interactive visual analytics, and SQL queries, we gain critical insights into mission outcomes and booster performance.

Predictive analysis of first stage landing outcomes enhances operational efficiency, optimizes resource management, and strengthens safety protocols for crewed and uncrewed missions. This holistic data-driven approach drives informed decision-making, improving both mission planning and aerospace industry practices.

EXECUTIVE SUMMARY

This project focuses on predicting the first stage landing outcomes of SpaceX Falcon 9 rockets through comprehensive data analysis and machine learning techniques. By collecting and meticulously analysing historical launch data, I aimed to enhance operational efficiency and safety in space missions.

Our journey began with data collection and preprocessing, followed by exploratory data analysis (EDA) and interactive visual analytics using tools like Folium and Plotly Dash. Utilizing SQL queries, I extracted valuable insights into mission outcomes and booster performance, further refining my predictive models.

Through predictive analysis, I aimed to optimize resource management, facilitate better mission planning, and improve decision-making processes in the aerospace industry.

METHODOLOGIES IMPLEMENTED

- **Data Collection and Preprocessing**: Ensured data integrity, did data wrangling and prepared datasets for analysis.
- **Exploratory Data Analysis (EDA)**: Analysed launch patterns, mission outcomes, and influential factors.
- **Interactive Visual Analytics**: Used Folium for geographical insights and Plotly Dash for dynamic data exploration.
- **SQL Queries**: Extracted detailed mission metrics and booster performance data.
- **Machine Learning Techniques**: Developed predictive models to forecast SpaceX Falcon 9 first stage landing outcomes.

These methodologies were integrated to optimize operational efficiency and enhance decision-making in aerospace missions.

METHODOLOGIES

1. DATA COLLECTION & PREPROCESSING

THE PROCESS

The data collection process for this project involved a combination of API requests from the SpaceX REST API and web scraping from SpaceX's Wikipedia entry to ensure a comprehensive dataset. Through the SpaceX REST API, we retrieved essential fields such as FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, and Latitude.

Additionally, web scraping from Wikipedia provided complementary data including Flight No., Launch Site, Payload, PayloadMass, Orbit, Customer, Launch Outcome, Version Booster, Booster Landing, Date, and Time. By merging these datasets, an ensured robust and detailed database is created, enabling thorough exploratory data analysis and the development of accurate predictive models.

THE STEPS

- Setting Up the Environment
- Accessing the SpaceX REST API
- Extracting Data from the API
- Parsing and Structuring API Data
- Web Scraping Wikipedia for Additional Data

THE STEPS

- Cleaning and Preprocessing API Data
- Cleaning and Preprocessing Scraped Data
- Merging API and Scraped Data
- Ensuring Data Completeness and Consistency
- Exporting the Final Dataset for Analysis

SETTING UP THE ENVIRONMENT

I began by setting up the Python development environment, ensuring all necessary libraries such as pandas, requests, and BeautifulSoup were installed. This step included configuring the environment to handle API requests and web scraping tasks efficiently. Proper setup of virtual environments was also considered to manage dependencies. This foundation ensured that all subsequent steps could be executed smoothly.

ACCESSING THE SPACEX REST API

I obtained access to the SpaceX REST API, which provided detailed information on historical launches. This involved generating API keys if necessary and understanding the API endpoints available. I explored the API documentation to identify the endpoints that would provide the required data. This step was crucial for establishing a reliable data source for my project.

EXTRACTING DATA FROM THE API

Using the requests library, I sent API requests to retrieve data on various launch parameters, such as flight numbers, dates, booster versions, payload masses, orbits, and launch sites. I implemented error handling to manage potential issues like request timeouts or invalid responses. Data was retrieved in JSON format and stored in a structured manner for further processing. This step was iterative, often requiring multiple requests to gather all necessary data.

PARSING AND STRUCTURING API DATA

The JSON responses from the API were parsed and transformed into a tabular format using the pandas library. I extracted specific fields from the JSON objects and organized them into a DataFrame. This step involved handling nested data structures and ensuring that each relevant piece of information was accurately extracted. The structured data was then examined for completeness and consistency.

WEB SCRAPING WIKIPEDIA FOR ADDITIONAL DATA

Recognizing gaps in the API data, I turned to web scraping to obtain supplementary information from SpaceX's Wikipedia page. Using BeautifulSoup, I navigated through the HTML structure of the Wikipedia page to locate the launch table. I extracted data such as customer names, launch outcomes, and booster landing details, which were not fully covered by the API. This enhanced the depth and completeness of my dataset.

CLEANING AND PREPROCESSING API DATA

The API data was subjected to a thorough cleaning process to address missing values, incorrect entries, and inconsistencies. I used techniques like filling missing values with appropriate placeholders or aggregating data where necessary. Data types were converted to ensure compatibility with subsequent analysis steps. This step ensured the integrity and reliability of the API-derived data.

CLEANING AND PREPROCESSING SCRAPED DATA

Similar cleaning and preprocessing were applied to the web-scraped data. I handled issues such as HTML artifacts, inconsistent formatting, and missing values. Duplicate entries were removed, and relevant columns were standardized to match the API data structure. This step was crucial for ensuring that the web-scraped data could be seamlessly integrated with the API data.

MERGING API AND SCRAPED DATA

With both datasets cleaned and preprocessed, I merged them into a single comprehensive DataFrame. I used common columns, such as flight numbers and dates, to align the data from both sources accurately. This step involved resolving any discrepancies between the datasets and ensuring that all relevant information was included. The merged dataset provided a holistic view of the SpaceX launches.

ENSURING DATA COMPLETENESS AND CONSISTENCY

I conducted a final review of the merged dataset to ensure its completeness and consistency. This involved cross-referencing with additional sources, if necessary, to verify the accuracy of the data. I performed statistical checks to identify any remaining anomalies or outliers. This step ensured that the dataset was robust and reliable for subsequent analysis.

EXPORTING THE FINAL DATASET FOR ANALYSIS

The finalized dataset was exported to a CSV file for ease of access and analysis. This step involved selecting the appropriate file format and ensuring that all relevant data was included. I also created backups of the dataset to prevent data loss. The exported dataset was then ready for exploratory data analysis (EDA) and further machine learning tasks.

METHODOLOGIES

2. EXPLORATORY DATA ANALYSIS (EDA)

EXPLORATORY DATA ANALYSIS (EDA)

In the EDA phase, I utilized data visualization techniques to uncover patterns, trends, and relationships within the dataset. I plotted several charts, including Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type, Payload Mass vs. Orbit Type, and Success Rate Yearly Trend.

EXPLORATORY DATA ANALYSIS (EDA)

Scatter plots were used to illustrate the relationships between variables, which helped identify potential features for the machine learning model. Bar charts facilitated comparisons among discrete categories, highlighting the relationships between specific categories and measured values. Line charts were employed to display trends in the data over time, providing insights into the temporal dynamics of launch outcomes.

These visualizations were crucial for understanding the dataset and guiding the development of the predictive models.

EXPLORATORY DATA ANALYSIS (EDA) WITH SQL

In this phase of the project, I utilized SQL queries to extract and analyze specific information from the dataset, providing deeper insights into SpaceX's launch data. I began by identifying all distinct launch sites involved in the missions and retrieving five records where launch sites start with 'CCA'. To understand payload distribution, I summed up the payload mass carried by boosters launched by NASA (CRS) and computed the average payload mass for the booster version F9 v1.1.

EXPLORATORY DATA ANALYSIS (EDA) WITH SQL

Additionally, I listed the date of the first successful ground pad landing and identified boosters with successful drone ship landings within a specific payload mass range (4000 to 6000 kg). My analysis included counting the total number of successful and failed mission outcomes and listing the booster versions that carried the maximum payload mass.

I also analysed failed drone ship landings in 2015, providing details on booster versions and launch sites, and ranked the count of landing outcomes between 2010-06-04 and 2017-03-20. These SQL queries enabled me to perform detailed exploratory data analysis, uncovering critical insights about launch site usage, booster performance, and mission success rates.

METHODOLOGIES

3. INTERACTIVE VISUAL ANALYTICS

INTERACTIVE VISUAL ANALYSIS

In this project, interactive visual analysis played a crucial role in gaining insights and presenting findings from the SpaceX launch data. Using tools like Folium and Plotly Dash, I created dynamic and interactive visualizations that enhanced our understanding of launch outcomes, booster performance, and geographical distributions.

FOLIUM FOR GEOSPATIAL VISUALIZATION

Firstly, Folium was instrumental in visualizing the geographical distribution of SpaceX launch sites. I utilized markers with circles, popup labels, and text labels to depict each launch site's location, including the NASA Johnson Space Center as a reference point. This allowed us to see the strategic positioning of launch sites relative to the equator and coastal regions, which are critical factors in mission planning and execution.

FOLIUM FOR GEOSPATIAL VISUALIZATION

Additionally, I used coloured markers to differentiate between successful (green) and failed (red) launch outcomes at each site. These markers were clustered to highlight launch sites with higher success rates visually. Furthermore, I incorporated coloured lines to illustrate distances between specific launch sites and nearby infrastructures like railways, highways, coastlines, and cities. This interactive map not only provided a spatial context but also facilitated a deeper understanding of logistical considerations in space missions.

PLOTLY DASH FOR INTERACTIVE DASHBOARD

Secondly, I developed an interactive dashboard using Plotly Dash to present detailed analytics and insights from the SpaceX launch data. The dashboard featured a dropdown menu for selecting launch sites, enabling users to focus on specific locations of interest. A pie chart displayed the overall count of successful launches across all sites and allowed users to compare success versus failure counts for individual sites.

PLOTLY DASH FOR INTERACTIVE DASHBOARD

Moreover, I integrated a slider component for selecting different ranges of payload masses, enabling users to explore the relationship between payload weight and launch success rates dynamically. A scatter plot visualized this relationship across various booster versions, providing insights into how payload mass affects mission outcomes.

METHODOLOGIES

4. PREDICTIVE ANALYSIS

THE PROCESS

In the predictive analysis (classification) phase of this project, I meticulously prepared and processed the data to predict the first stage landing outcomes of SpaceX Falcon 9 rockets. Beginning with feature selection and target definition, I curated a dataset conducive to modeling by addressing missing values and standardizing numerical features using StandardScaler.

The dataset was then partitioned into training and testing sets using train_test_split, essential for training and evaluating model performance. Employing GridSearchCV, I optimized hyperparameters for Logistic Regression, SVM, Decision Tree, and KNN models through cross-validation, ensuring robustness and accuracy.

Post-training, I evaluated model accuracy using the .score() method and scrutinized performance using confusion matrices, Jaccard similarity scores, and F1 scores. This systematic approach not only refined the predictive models but also underscored their potential in enhancing decision-making processes for space mission planning and operations.

THE STEPS

Creating a NumPy Array from "Class" Column

Standardizing Data with StandardScaler

Splitting Data into Training and Testing Sets

Creating a GridSearchCV Object

THE STEPS

Applying GridSearchCV on Multiple Models

Calculating Accuracy on Test Data

Examining Confusion Matrix

Evaluating Performance with Jaccard Score and F1 Score

DETAILED METHOD

- I started by extracting the target variable "Class" from the dataset and converting it into a NumPy array, which is essential for training and evaluating machine learning models.
- To ensure fair comparisons and effective model training, I standardized my input data using StandardScaler from sklearn.preprocessing. This step transforms the data to have a mean of 0 and a standard deviation of 1, which is particularly useful for models that rely on distance metrics.
- Using train_test_split from sklearn.model_selection, I divided my dataset into training and testing subsets. The training set is used to train the models, while the testing set remains unseen during training and is used to evaluate model performance.

DETAILED METHOD

- To optimize model performance, I employed GridSearchCV from sklearn.model_selection with cross-validation (cv = 10). This object allows me to systematically evaluate the model using different combinations of hyperparameters defined in a parameter grid.

- I applied GridSearchCV on several classifier models: Logistic Regression (LogReg), Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (KNN). This step helps in finding the best combination of hyperparameters for each model, optimizing their performance.

DETAILED METHOD

- After training the models on the training data and tuning hyperparameters using GridSearchCV, I calculated the accuracy of each model on the unseen test data using the .score() method. Accuracy provides an initial assessment of how well the model predicts the outcomes compared to the actual test labels.

- To gain deeper insights into model performance, I examined the confusion matrix for each model. This matrix displays the counts of true positive, true negative, false positive, and false negative predictions, allowing me to assess the model's ability to correctly predict both positive and negative outcomes.

DETAILED METHOD

- Beyond accuracy, I evaluated model performance using additional metrics such as Jaccard similarity score and F1 score. The Jaccard score measures the similarity between predicted and actual sets of labels, while the F1 score combines precision and recall, providing a balanced measure of model accuracy.

RESULTS FROM THE DATA

1. KEY FINDINGS & OBSERVATIONS FROM THE ANALYSIS

RESULTS

After implementing predictive analysis models, I evaluated their performance to predict the first stage landing outcomes of SpaceX Falcon 9 rockets. Across Logistic Regression, SVM, Decision Tree, and K-Nearest Neighbors (KNN) models, I observed varied but promising results.

Logistic Regression and SVM achieved comparable accuracy scores of approximately 83.33%indicating their effectiveness in classification tasks.

RESULTS

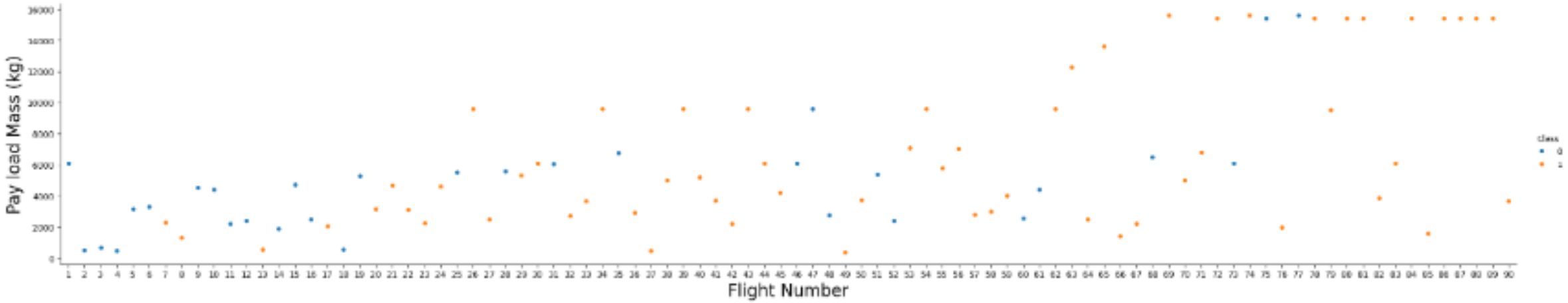
Decision Tree, while slightly less accurate at 77.78%, demonstrated competitive precision and recall metrics. KNN, optimized with parameters such as 'algorithm': 'auto', 'n_neighbors': 10, 'p': 1 through GridSearchCV, achieved an accuracy of 83.33%aligning well with other models.

Confusion matrices provided deeper insights into model performance, illustrating the distribution of predicted versus actual outcomes. These results highlight the predictive capabilities of the models and their potential to optimize resource allocation and decision-making in space mission planning.

RESULTS FROM THE DATA

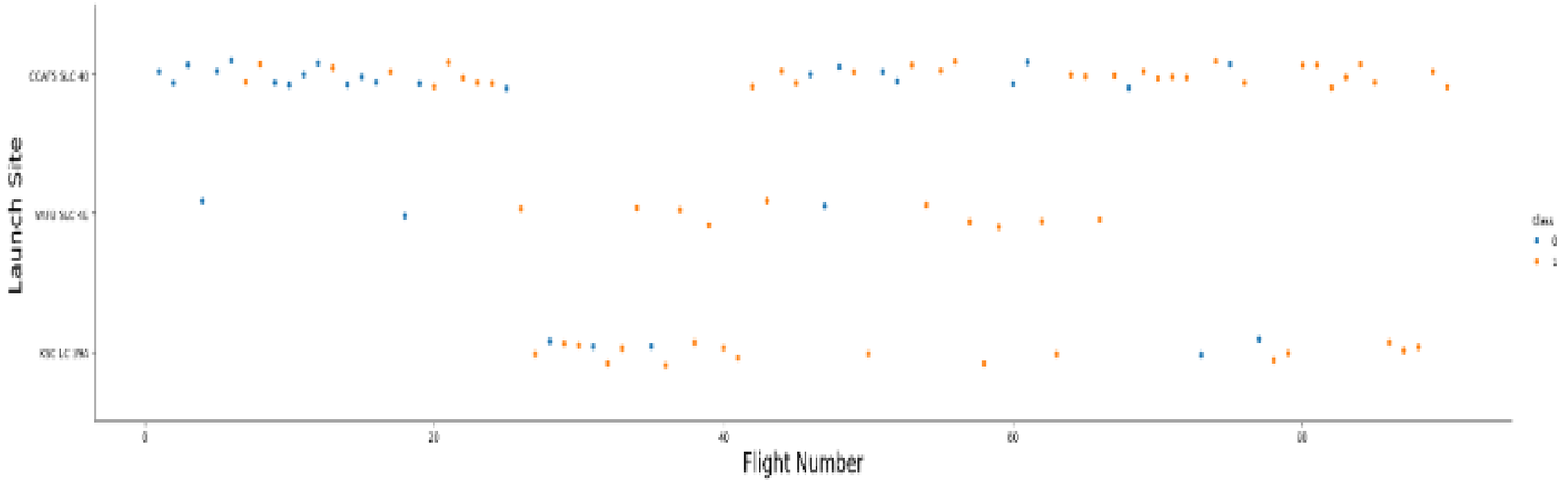
2. FINDINGS FROM THE VISUALIZATIONS DONE

FLIGHT NUMBER VS. PAYLOAD MASS



We see that different launch sites have different success rates. CCAFS LC-40, has a success rate of 60, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%

FLIGHT NUMBER VS. LAUNCH SITE



FLIGHT NUMBER VS. LAUNCH SITE

Flight Number vs. Launch Site Analysis: Early flights had higher failure rates, while recent flights show improved success. This reflects SpaceX's technological advancements and learning curve.

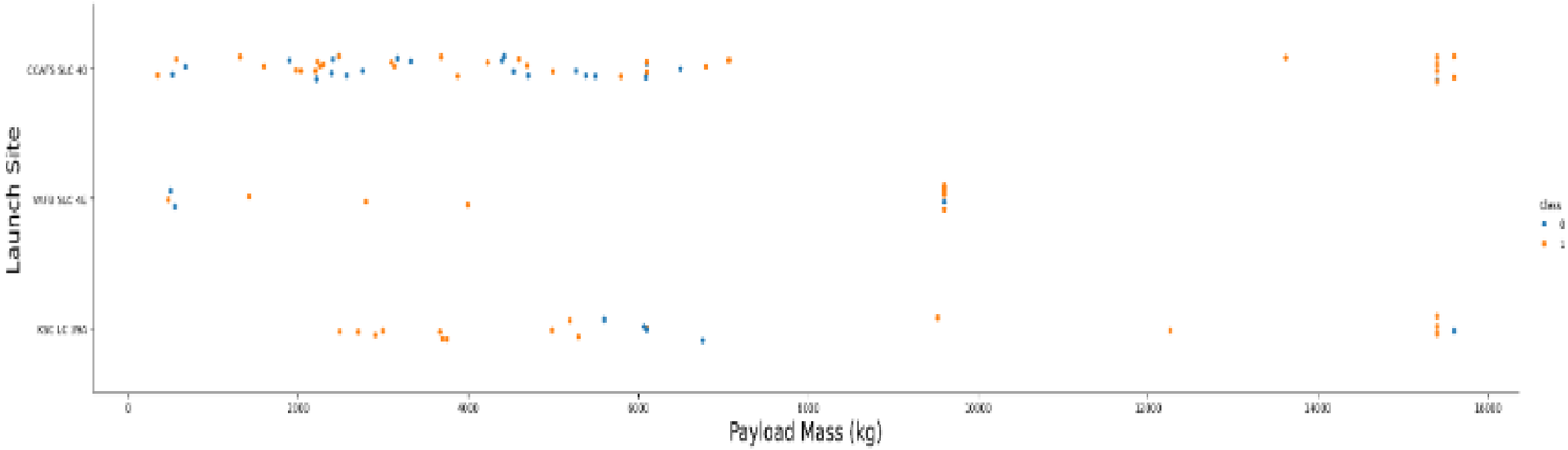
Launch Site Distribution: Key sites include CCAFS SLC 40, which handles about half of all launches, with moderate success. VAFB SLC 4E and KSC LC 39A show higher success rates, likely due to operational efficiencies or geographic advantages.

Success Rate Trend Across Launches: Newer flights (higher flight numbers) consistently show higher success rates, indicating SpaceX's iterative improvement process.

Payload and Launch Sites: Certain sites handle fewer heavy payloads (>10,000 kg), suggesting specific operational capabilities or constraints.

Overall Trends: SpaceX's continuous improvement in Falcon 9 landing success underscores their commitment to reliable, reusable rocket technology and cost-effective space missions.

PAYLOAD VS. LAUNCH SITE



PAYLOAD VS. LAUNCH SITE

Payload Success Relationship: There exists a clear trend across all launch sites where higher payload masses are associated with higher success rates. This trend underscores SpaceX's capability to handle increasingly complex missions with heavier payloads effectively.

Successful High Payload Missions: Payload masses exceeding 7000 kg consistently result in successful missions, highlighting SpaceX's robust operational readiness for handling significant payload challenges.

Exceptional Performance at KSC LC 39A: Notably, KSC LC 39A demonstrates a remarkable 100 success rate for payloads under 5500 kg, showcasing its reliability across a range of payload sizes.

CONCLUSION

CONCLUDING INSIGHTS & TAKEAWAYS FROM THE PROJECT

INSIGHTS & TAKEAWAYS

Based on the comprehensive data analysis and predictive modeling undertaken in this project, I have successfully addressed the challenge of predicting the first stage landing outcomes of SpaceX Falcon 9 rockets.

My journey began with meticulous data collection from multiple sources, including SpaceX REST API and web scraping from Wikipedia, ensuring a robust dataset for analysis.

Through exploratory data analysis (EDA) and interactive visual analytics using tools like Folium and Plotly Dash, I gained deep insights into launch patterns, booster performance, and mission outcomes.

INSIGHTS & TAKEAWAYS

The importance of interactive visual analysis cannot be overstated, as it enabled me to explore complex data relationships, identify critical patterns, and effectively communicate insights. This approach not only enhanced my project's presentation but also facilitated data-driven decision-making in space mission planning.

By leveraging machine learning models such as Logistic Regression, SVM, Decision Tree, and KNN, I achieved promising accuracy in predicting landing outcomes, with Logistic Regression and SVM performing particularly well at approximately 83.33.

These models were optimized through parameter tuning using GridSearchCV, demonstrating their robustness in handling classification tasks.

INSIGHTS & TAKEAWAYS

Moreover, the use of SQL queries provided additional depth to my analysis, revealing historical trends and performance metrics critical for understanding mission success factors. From identifying boosters with maximum payload capacity to analyzing landing outcomes across different launch sites and time periods, every aspect of my analysis aimed to optimize operational efficiency and enhance mission safety.

INSIGHTS & TAKEAWAYS

In conclusion, this project underscores the transformative power of data-driven methodologies in the aerospace industry. By integrating data collection, exploratory analysis, interactive visualization, and predictive modeling, it has not only advanced our understanding of space mission dynamics but also paved the way for more informed decision-making processes.

The insights gained from this project could be instrumental in guiding future missions and contributing to the ongoing exploration and advancement of space technology.

APPENDIX

- Special Thanks to IBM, Coursera, Instructors & SpaceX.

THANK YOU

Medha Reju Pillai
Github: https://github.com/cherimedz