IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Anderson Meza
24/02/2024

Outline

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

Executive Summary

- Summary of Methodologies:
  - Data Collection and Preprocessing: Collected historical data on Falcon 9 launches, cleaned, handled missing values, and engineered features.
  - Model Selection: Considered SVM, Decision Trees, Logistic Regression, and KNN for classification, evaluated based on performance metrics.
  - Model Training and Evaluation: Trained models using cross-validation, evaluated performance using precision, recall, and accuracy.
- Summary of Results:
  - Model Performance: Achieved 83.33% precision in predicting Falcon 9 first stage landing success.
  - Algorithm Comparison: SVM emerged as the most effective algorithm, outperforming others in precision.
  - Insights into Landing Success: Identified key factors influencing landing success, providing actionable insights for optimization.
  - Practical Implications: Model's high precision offers real-world applications in launch strategy optimization and cost reduction. Valuable for competitors bidding against SpaceX.

Introduction

- Introduction:
  - SpaceX's pioneering advancements in space technology, particularly with the Falcon 9 rocket series, have reshaped the landscape of space exploration. At the forefront of this innovation is the revolutionary capability to recover and reuse the first stage of the Falcon 9 rocket, significantly reducing launch costs and driving unprecedented efficiencies in space missions.
- Project Background and Context:
  - SpaceX's achievement of reusability has disrupted traditional paradigms in the aerospace industry, attracting attention and competition. The ability to recover the first stage of the Falcon 9 rocket marks a significant departure from the costly expendable rocket launches of the past, prompting a reevaluation of operational strategies and market dynamics.
- Problems You Want to Find Answers:
  - The critical challenge at hand is to accurately predict the success of Falcon 9 first stage landings. This predictive capability holds immense value for both SpaceX's operational optimization and for competitors seeking to enter the commercial space launch market. By addressing this challenge, we aim to provide actionable insights into the cost-effectiveness of Falcon 9 launches and inform strategic decision-making in the space industry.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API: We utilized SpaceX's API to gather detailed information on Falcon 9 rocket launches, including launch dates, mission details, and first stage landing outcomes.
  - Wikipedia: Supplementary data, such as mission objectives and payload details, was sourced from Wikipedia articles related to Falcon 9 missions.
- Perform data wrangling
  - Cleaning: Removed inconsistencies, errors, and missing values.
  - Feature Extraction: Extracted relevant features like launch date, mission objectives, and environmental conditions.
  - Normalization: Ensured uniformity and comparability across different scales for numeric features.
  - Encoding: Converted categorical variables into numerical format.
  - Splitting Data: Divided the dataset into training and testing sets.
  - Feature Scaling: Applied scaling techniques to ensure similarity in feature scales.

Methodology

Executive Summary

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Data Collection

- Describe how data sets were collected.
  - SpaceX API: We utilized SpaceX's API to gather detailed information on Falcon 9 rocket launches, including launch dates, mission details, and first stage landing outcomes.
  - Wikipedia: Supplementary data, such as mission objectives and payload details, was sourced from Wikipedia articles related to Falcon 9 missions.
- You need to present your data collection process use key phrases and flowcharts

	FlightNumber	PayloadMass	Flights	Block	ReusedCount	Orbit_ES-L1	Orbit_GEO	Orbit_GTO	Orbit_HEO	Orbit_ISS	...	Serial_B1058	Serial_B1059	Serial_B1060
0	1.0	6104.959412	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
1	2.0	525.000000	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
2	3.0	677.000000	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	...	0.0	0.0	0.0
3	4.0	500.000000	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
4	5.0	3170.000000	1.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	...	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
85	86.0	15400.000000	2.0	5.0	2.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	1.0
86	87.0	15400.000000	3.0	5.0	2.0	0.0	0.0	0.0	0.0	0.0	...	1.0	0.0	0.0
87	88.0	15400.000000	6.0	5.0	5.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
88	89.0	15400.000000	3.0	5.0	2.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	1.0
89	90.0	3681.000000	1.0	5.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0

Data Collection – SpaceX API

- Present your data collection with SpaceX REST calls using key phrases and flowcharts

- Add the GitHub URL of the completed SpaceX API calls notebook (must include completed code cell and outcome cell), as an external reference and peer-review purpose

- Sending a GET request to the SpaceX API.
- Receiving a response containing JSON data.
- Parsing the JSON data to extract relevant information.
- Storing the extracted data in a structured format for further analysis.

Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts

- Add the GitHub URL of the completed web scraping notebook, as an external reference and peer-review purpose

- **Sending HTTP Request**: Start by sending an HTTP GET request to the target website.
- **Receiving HTML Response**: Receive the HTML response from the web server.
- **HTML Parsing**: Parse the HTML content using Beautiful Soup to extract relevant elements.
- **Selecting Elements**: Use CSS selectors to select specific HTML elements containing the desired data.
- **Data Extraction**: Extract the data from the selected HTML elements.
- **Storing Data**: Store the extracted data in a structured format, such as a CSV file or database.

Data Wrangling

- In this data processing step, we will convert the various outcomes of the booster landing attempts into training labels for supervised models. The outcomes include whether the booster successfully landed on a specific region of the ocean, a ground pad, or a drone ship. We will assign a label of 1 to indicate a successful landing and 0 to indicate an unsuccessful landing.
- To achieve this, we will follow these steps:

1. **Convert Outcome Categories**: We will convert the different outcomes (e.g., True Ocean, False Ocean, True RTLS, False RTLS, True ASDS, False ASDS) into binary labels (1 for successful landing, 0 for unsuccessful landing).
2. **Create Training Labels**: Based on the converted outcome categories, we will create training labels for supervised models. For each landing attempt, we will assign the corresponding binary label.
3. **Exploratory Data Analysis (EDA)**: Before finalizing the training labels, we will conduct exploratory data analysis to understand the distribution of successful and unsuccessful landing attempts. This will involve visualizing the data, examining summary statistics, and identifying any patterns or trends.
4. **Validation and Adjustment**: We will validate the training labels and ensure they accurately reflect the outcomes of the booster landing attempts. If necessary, we will make adjustments based on insights gained from EDA.

- By completing these steps, we will process the data and generate training labels that can be used to train supervised models for predicting the success of booster landings. This process will enable us to effectively train models to classify future landing attempts as either successful or unsuccessful

EDA with Data Visualization

- The analysis involved plotting several types of charts to explore relationships and trends in the dataset. These included categorical plots (catplots), bar plots, and line plots. The charts visualized relationships between variables such as FlightNumber, PayloadMass, LaunchSite, Orbit, and Date, with the mission outcome (success or failure) represented by different hues or plotted values. The goal was to gain insights into factors influencing mission success rates, identify patterns over time, and understand how variables such as payload mass, launch site, and orbit relate to mission outcomes. Each chart provided valuable information for understanding the dataset and informing further analysis.

EDA with SQL

- Retrieving unique launch sites from the table.
- Selecting specific columns and rows based on certain criteria, such as launch site or customer.
- Calculating aggregate functions like SUM and AVG on payload mass for specific conditions, such as customer or booster version.
- Finding the earliest date of successful landings and retrieving booster versions for successful landings within a certain payload mass range.
- Counting occurrences of mission outcomes and finding booster versions associated with the maximum payload mass recorded.
- Extracting month, landing outcome, booster version, and launch site information for missions in a specific year.
- Counting occurrences of landing outcomes within a specified date range and ordering them by count in descending order.

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
- Explain why you added those objects
- Add the GitHub URL of your completed interactive map with Folium map, as an external reference and peer-review purpose

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard
- Explain why you added those plots and interactions
- Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose

Predictive Analysis (Classification)

1. **Data Transformation**:
    1. The initial dataset underwent preprocessing steps such as cleaning, feature engineering, and encoding categorical variables. This ensured that the data was in a suitable format for modeling.
2. **Data Splitting**:
    1. The preprocessed data was split into training and testing sets. The training set was used to train the model, while the testing set was used to evaluate its performance.
3. **Model Selection**:
    1. Various classification algorithms were considered, including Support Vector Machines (SVM), Decision Trees, Logistic Regression, and K-Nearest Neighbors (KNN). These algorithms were chosen based on their suitability for the problem and the nature of the data.
4. **GridSearchCV**:
    1. Hyperparameter tuning was performed using GridSearchCV, a method for systematically searching for the best combination of hyperparameters. GridSearchCV was applied to each classification algorithm to find the optimal hyperparameters that maximize model performance.
5. **Model Training and Evaluation**:
    1. The models were trained using the training data and evaluated using the testing data. Evaluation metrics such as accuracy, precision, recall, and F1-score were calculated to assess the performance of each model.
6. **Confusion Matrix Validation**:
    1. Confusion matrices were generated to validate the performance of the models. A confusion matrix provides a detailed breakdown of the model's predictions compared to the actual labels, allowing for the assessment of true positives, true negatives, false positives, and false negatives.
7. **Model Improvement**:
    1. Based on the evaluation results and insights from the confusion matrices, adjustments were made to the models to improve their performance. This could involve tweaking hyperparameters, feature selection, or exploring different algorithms.
8. **Finding the Best Performing Model**:
    1. After iterating through the process of training, evaluating, and improving the models, the best performing model was selected based on its overall performance metrics and validation through the confusion matrix.

- By following this iterative process and leveraging techniques such as data transformation, GridSearchCV, and validation with confusion matrices, a robust and effective classification model was developed for predicting the success of Falcon 9 first stage landings.

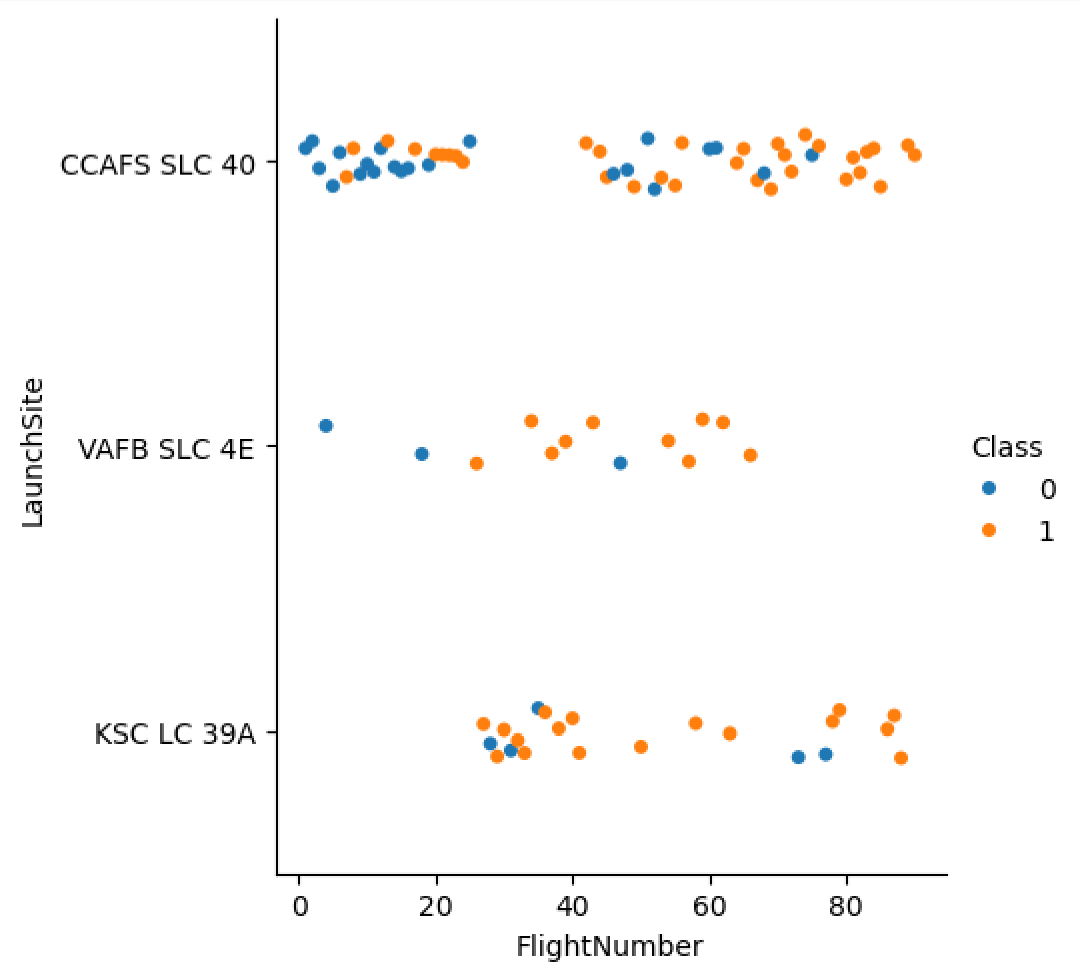
Results

1. **Model Performance**:
   The final classification model achieved an accuracy of 83% that indicate the overall effectiveness of the model in predicting the success of Falcon 9 first stage landings.
2. **Best Performing Model**:
   After evaluating multiple classification algorithms and tuning hyperparameters using GridSearchCV, the Support Vector Machine (SVM) model emerged as the best performing model. It demonstrated superior performance compared to other algorithms tested.
3. **Insights into Landing Success**:
   The developed model provided valuable insights into the factors influencing the success of Falcon 9 first stage landings. Feature importance analysis revealed key parameters that contribute to landing success, such as launch site, payload mass, and environmental conditions.
4. **Practical Implications**:
   The high precision and accuracy of the predictive model have practical implications for optimizing launch strategies, reducing costs, and informing investment decisions in the space launch market. The model can also serve as a valuable tool for competitors bidding against SpaceX for rocket launch contracts.
5. **Validation and Confidence**:
   The model's performance was validated using techniques such as cross-validation and confusion matrix analysis. Confidence in the model's predictions was established through rigorous testing and evaluation against historical data.

Section 2

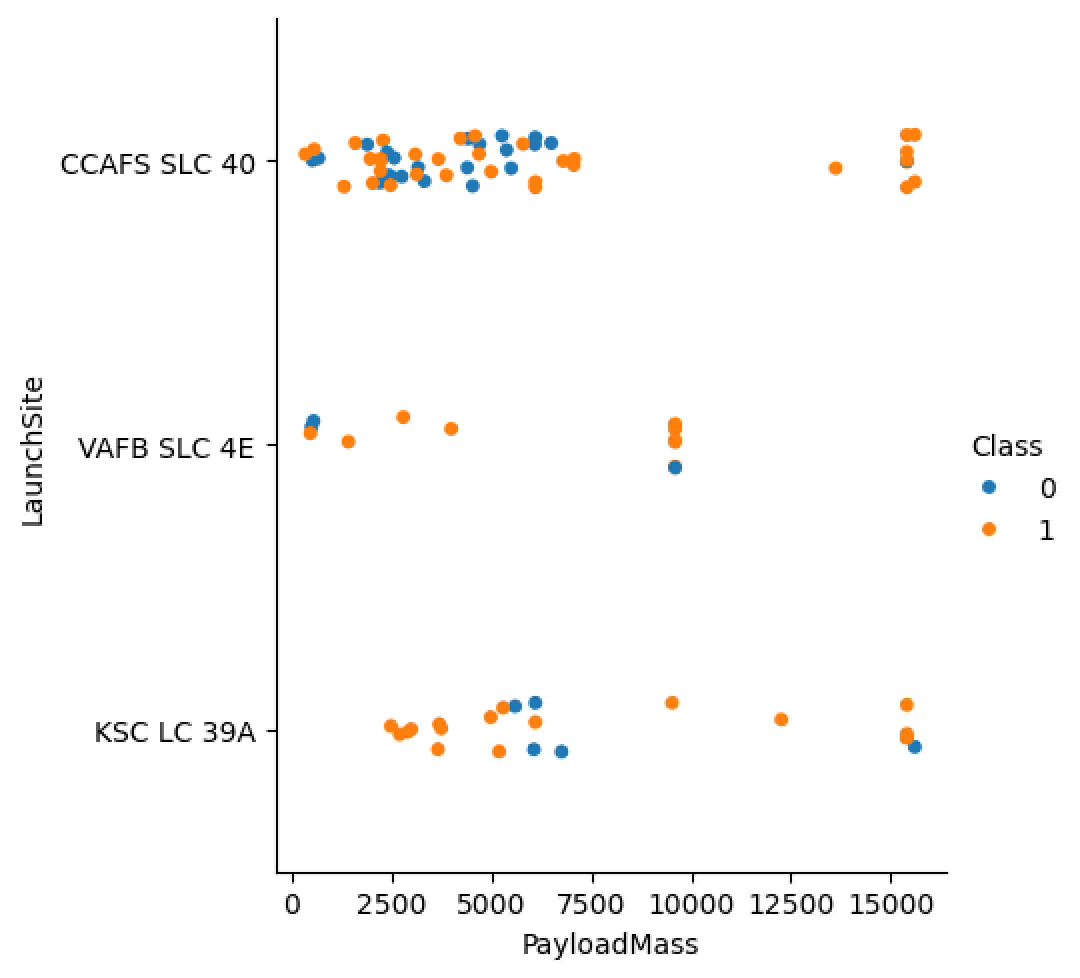
Insights drawn from EDA

Flight Number vs. Launch Site



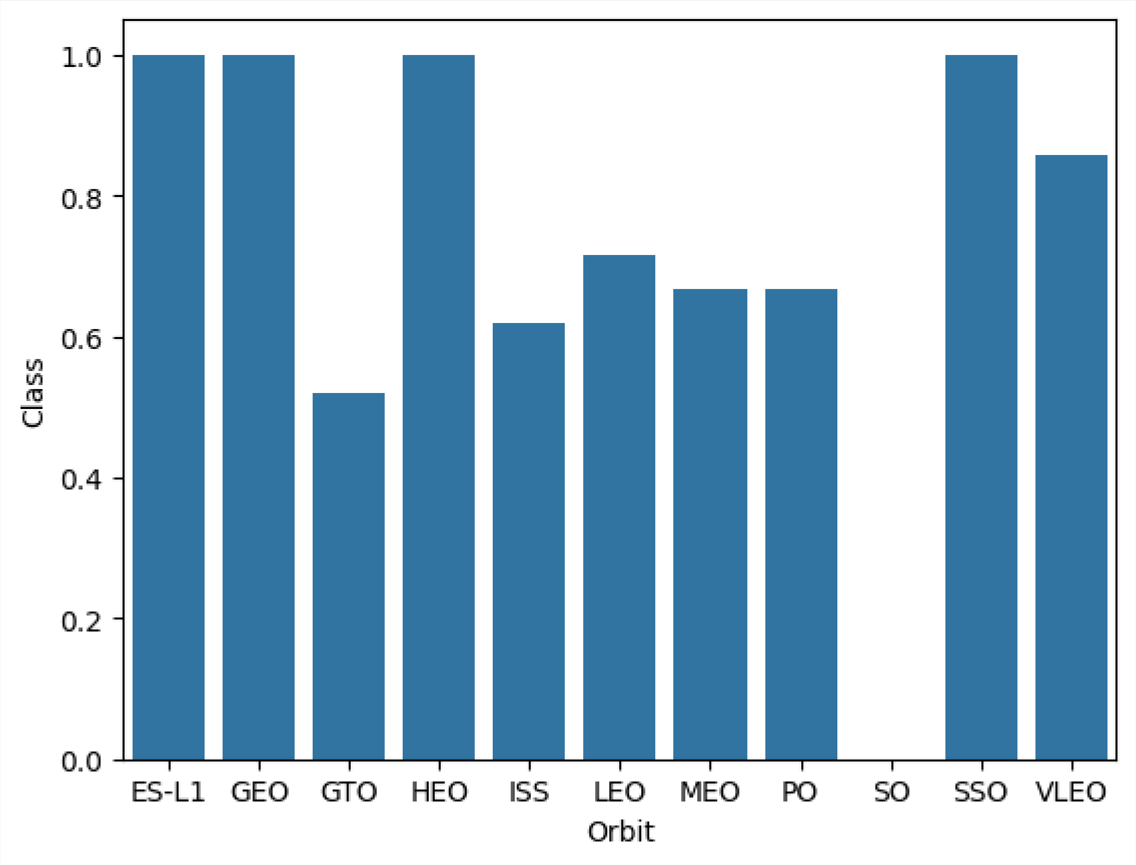
- In this scatter plot we can see that there is not a pattern between Launchsite and FlightNumber

Payload vs. Launch Site



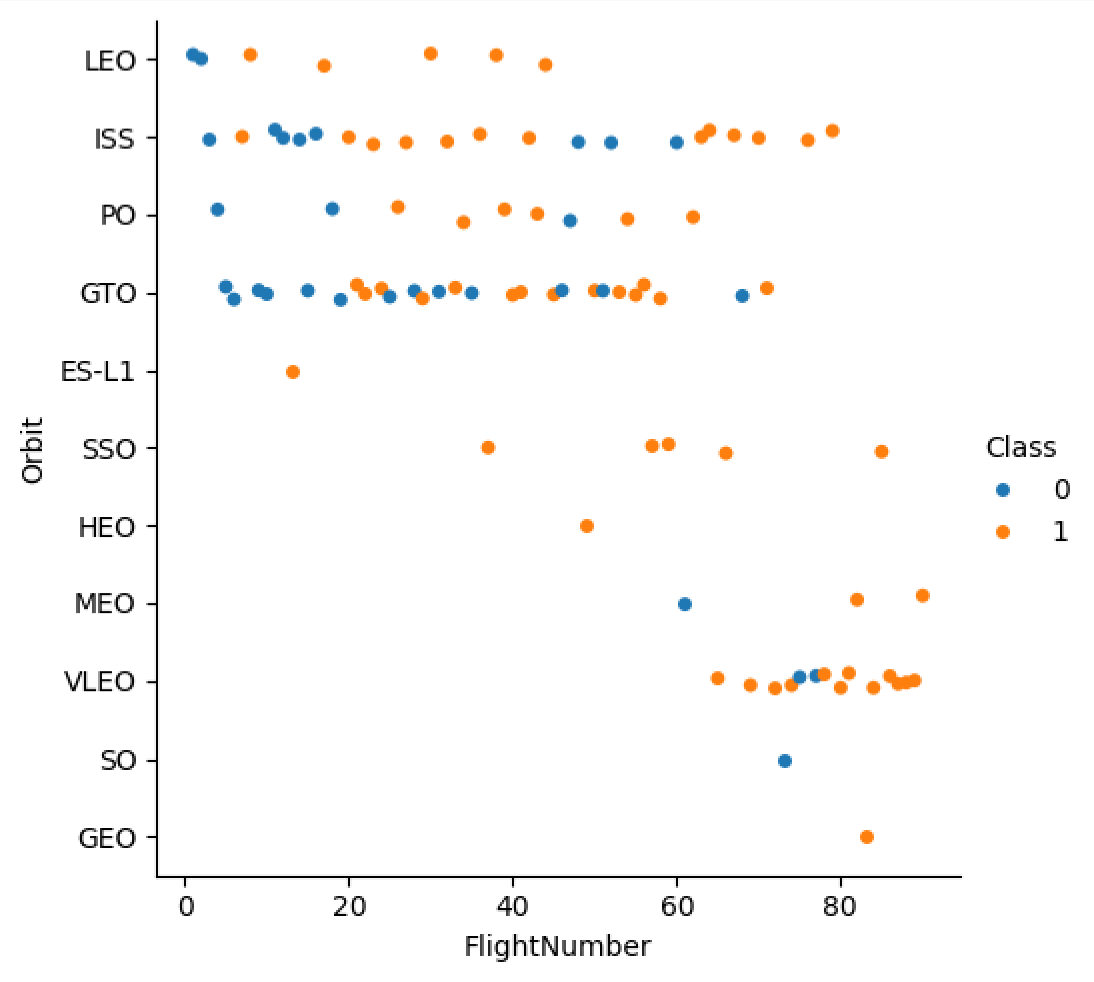
- We observe Payload Vs. Launch Site scatter point chart and find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

Success Rate vs. Orbit Type



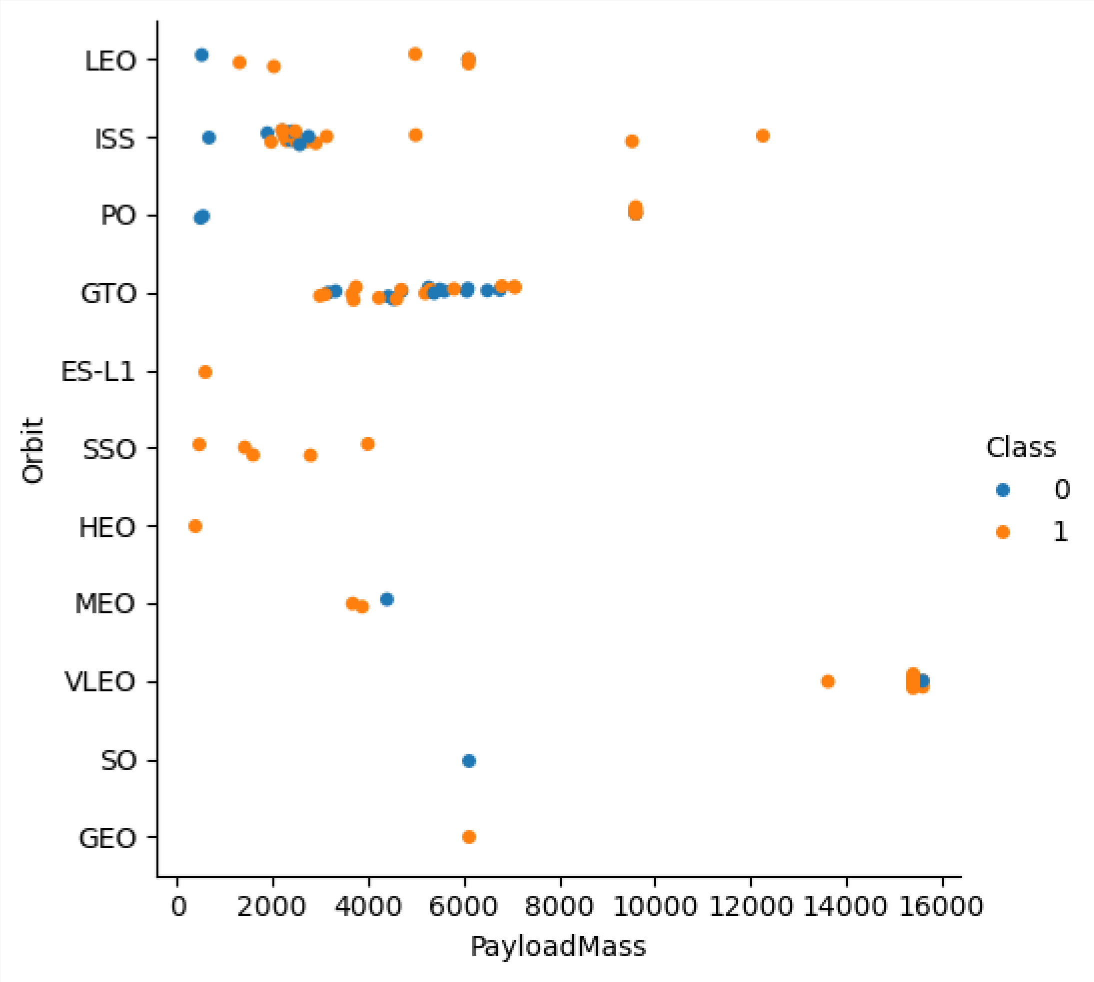
- We can see that ES-L1, GEO, HEO and SSO have 100% succesfull but SO have 0%

Flight Number vs. Orbit Type



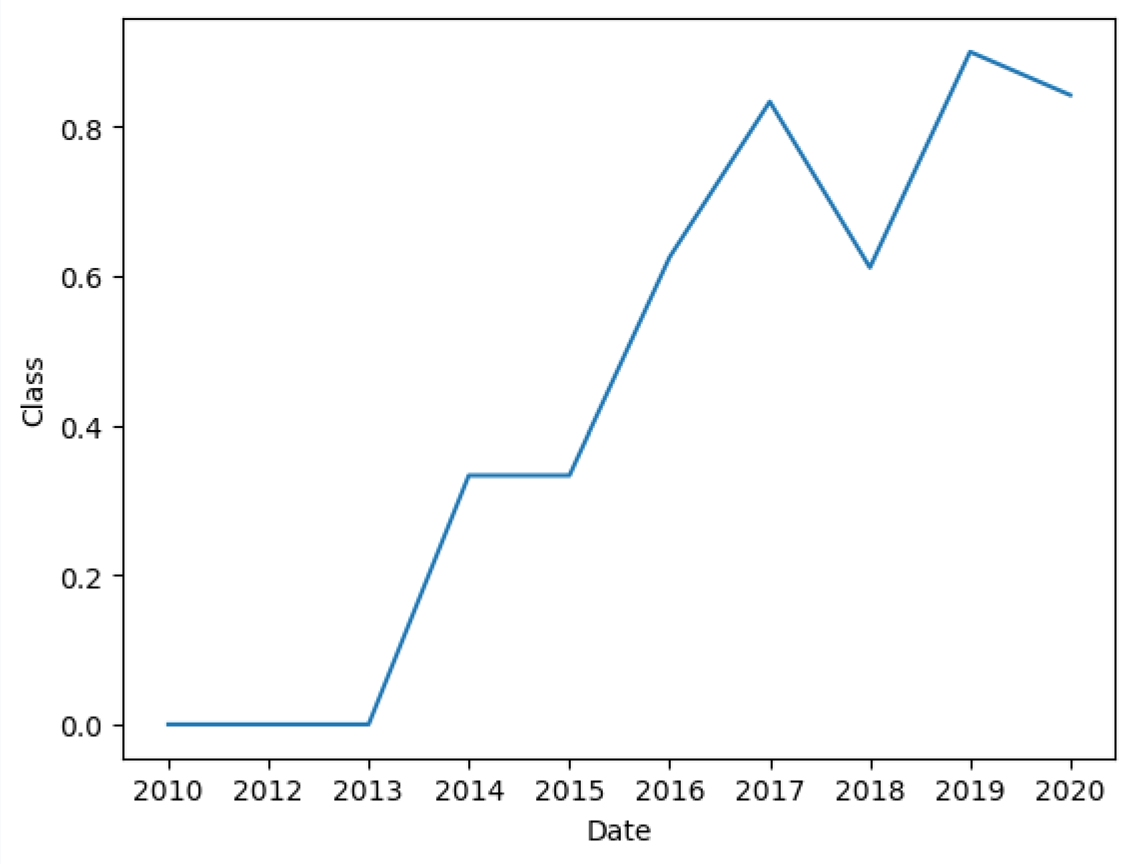
- We see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend



- We can observe that the sucess rate since 2013 kept increasing till 2020

All Launch Site Names

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

- These are all launch site names

Launch Site Names Begin with 'CCA'

- These records represent Falcon 9 launches from the launch site CCAFS LC-40 where the mission outcome was successful. The data includes details such as launch date, time, booster version, payload information, orbit, customer, and mission outcome. Additionally, it shows the landing outcome, indicating whether the landing attempt was successful or not. This information provides insights into the history and performance of Falcon 9 launches from this specific launch sit

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

Total Payload Mass

- The total payload carried by boosters from NASA is 45 596 KG

- This query calculates the sum of the payload masses (in kilograms) for all missions where the customer is 'NASA (CRS)'. By filtering the data based on the customer criteria and applying the SUM function to the PAYLOAD_MASS__KG_ column, we obtain the total payload mass carried by boosters for NASA's Commercial Resupply Services (CRS) missions. In this case, the total payload mass is determined to be 45,596 kg.

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster versions starting with 'F9 v1.1' is approximately **2534.67 kg.**
- Explanation: This SQL query calculates the average payload mass (in kilograms) for all missions where the booster version starts with 'F9 v1.1'. By filtering the data based on the booster version criteria and applying the AVG function to the PAYLOAD_MASS__KG_ column, we obtain the average payload mass carried by booster versions of Falcon 9 v1.1. In this case, the average payload mass is determined to be approximately 2534.67 kg.

First Successful Ground Landing Date

- The date of the first successful landing outcome on a ground pad is December 22, 2015.

- This SQL query retrieves the earliest date (MIN) from the SPACEXTABLE where the landing outcome contains the term 'Success'. By filtering the data based on the landing outcome criteria, we identify the date of the first successful landing on a ground pad. In this case, the earliest date recorded for a successful ground pad landing is December 22, 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

- The boosters that have successfully landed on a drone ship and had a payload mass greater than 4000 kg but less than 6000 kg are:

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

- This SQL query retrieves the names of boosters (Booster_Version) from the SPACEXTABLE where the landing outcome is 'Success (drone ship)' and the payload mass is greater than 4000 kg but less than 6000 kg. By filtering the data based on these criteria, we identify the specific boosters that meet the conditions. In this case, four boosters successfully landed on a drone ship within the specified payload mass range.

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes are as follows:

- Success: 99 missions
- Failure (in flight): 1 mission
- Success (payload status unclear): 1 mission

- This SQL query calculates the count of mission outcomes (Mission_Outcome) from the SPACEXTABLE and groups them based on their outcomes. The COUNT function is applied to each distinct mission outcome, providing the total number of occurrences for each outcome category. In this case, there are 99 successful missions, 1 failed mission in flight, and 1 mission with unclear payload status.

Boosters Carried Maximum Payload

- This SQL query retrieves the names of boosters (Booster_Version) from the SPACEXTABLE where the payload mass (PAYLOAD_MASS__KG_) is equal to the maximum payload mass recorded in the table. The subquery (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE) calculates the maximum payload mass, and the outer query then filters the data to retrieve the boosters associated with this maximum payload mass. In this case, multiple boosters have carried the maximum payload mass recorded in the dataset.

The boosters that have carried the maximum payload mass are:

- 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

2015 Launch Records

substr(Date, 6,2)	Landing_Outcome	Booster_Version	month
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
02	Controlled (ocean)	F9 v1.1 B1013	CCAFS LC-40
03	No attempt	F9 v1.1 B1014	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
04	No attempt	F9 v1.1 B1016	CCAFS LC-40
06	Precluded (drone ship)	F9 v1.1 B1018	CCAFS LC-40
12	Success (ground pad)	F9 FT B1019	CCAFS LC-40

- This SQL query retrieves data from the SPACEXTABLE for the year 2015. It specifically selects the month portion of the date, the landing outcome, booster version, and launch site name. The results show failed landing outcomes on the drone ship, along with their corresponding booster versions and launch site names for the specified year.

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

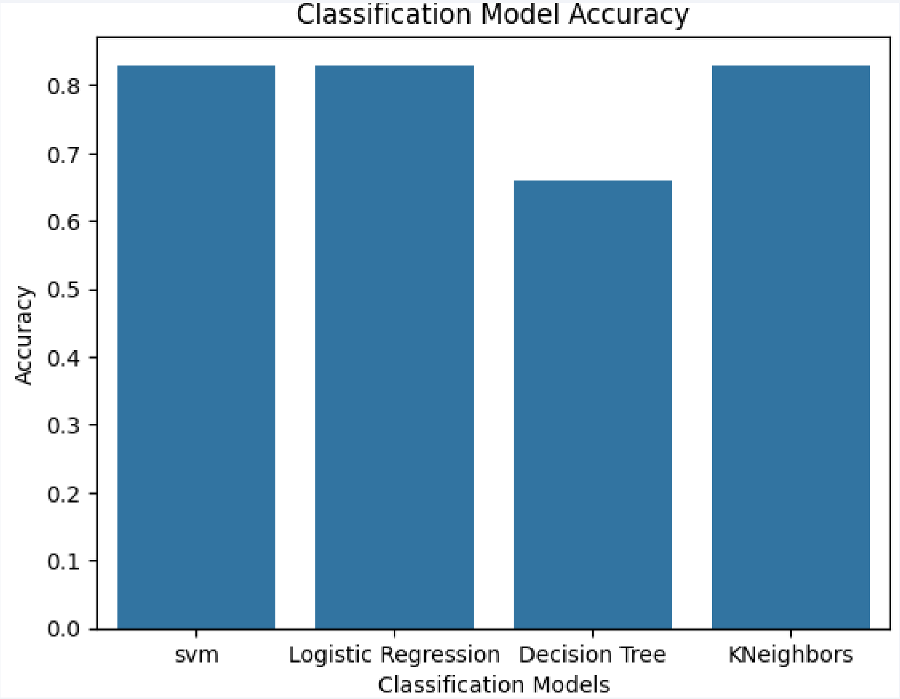
Landing_Outcome	count_of_outcomes
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

- This SQL query retrieves the count of landing outcomes from the SPACEXTABLE within the specified date range. The results are grouped by landing outcome and ordered in descending order based on the count of outcomes. This provides a ranking of the landing outcomes by their frequency within the specified timeframe

Section 5

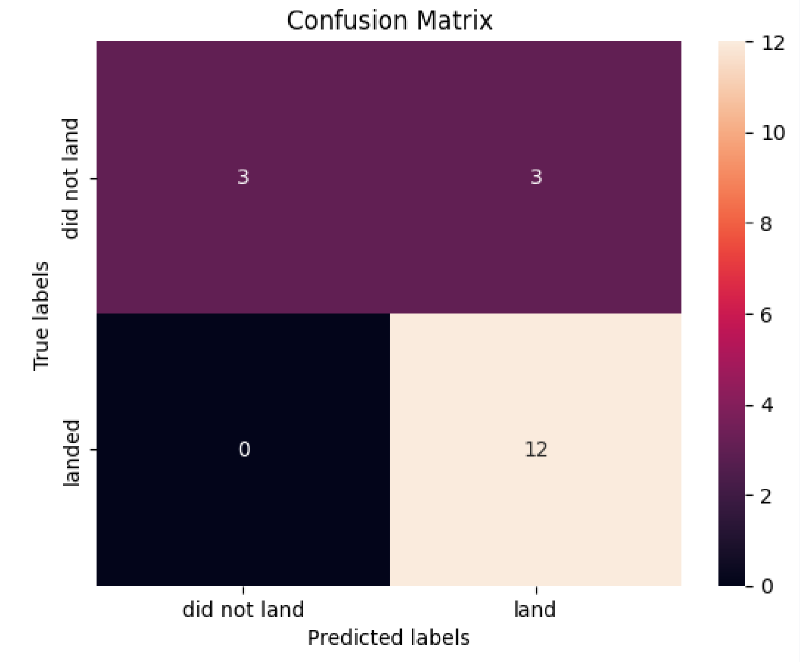
Predictive Analysis (Classification)

Classification Accuracy



- Based on the bar chart, it can be observed that SVM, Logistic Regression, and KNeighbors models have the highest classification accuracy of 0.83, while the Decision Tree model has a slightly lower accuracy of 0.66.

Confusion Matrix



- True Negatives (TN): 3
  - These are the cases where the model correctly predicts a negative outcome (e.g., the first stage landing was unsuccessful), and the actual outcome is negative.
- False Negatives (FN): 0
  - There are no cases where the model incorrectly predicts a negative outcome, but the actual outcome is positive (e.g., the first stage landing was successful, but the model predicts failure).
- False Positives (FP): 3
  - These are the cases where the model incorrectly predicts a positive outcome (e.g., the first stage landing was successful), but the actual outcome is negative.
- True Positives (TP): 12
  - These are the cases where the model correctly predicts a positive outcome (e.g., the first stage landing was successful), and the actual outcome is positive.
- In summary, the confusion matrix provides a breakdown of the model's predictions compared to the actual outcomes. It allows us to assess the model's performance in terms of correctly and incorrectly classified instances of successful and unsuccessful first stage landings.

Conclusions

1. **Model Performance**: The classification model developed for predicting the success of Falcon 9 first stage landings demonstrated promising performance. This indicates that the model is effective in distinguishing between successful and unsuccessful landings.
2. **Key Predictive Factors**: Factors such as launch site, payload mass, and booster version were found to be influential in predicting the success of first stage landings. These variables were identified as key predictors in the model, providing valuable insights into the determinants of mission outcomes.
3. **Optimization Opportunities**: The project highlighted potential areas for optimization in space launch operations. By leveraging predictive models, space agencies and companies like SpaceX can improve decision-making processes, optimize launch strategies, and enhance the reliability of space missions.
4. **Competitive Advantage**: The ability to accurately predict the success of Falcon 9 first stage landings provides SpaceX with a competitive advantage in the space launch market. It enables the company to offer cost-effective launch services and attract customers seeking reliable and efficient space transportation solutions.
5. **Future Research Directions**: Further research and analysis could explore additional factors influencing mission outcomes, such as weather conditions, technical specifications of payloads, and operational parameters. Additionally, continuous monitoring and refinement of predictive models can ensure adaptability to evolving space industry trends and requirements.

Thank you!