SpaceX Falcon 9 Rocket Recovery Prediction

This Coursera Data Science Project focuses on leveraging machine learning techniques to predict the success of Stage 1 recovery for the SpaceX Falcon 9 rocket. The ability to accurately forecast the recovery of this critical rocket component has significant implications for SpaceX's operations, enabling the company to optimize their launch procedures and reduce costs.

The project began with the collection of data from public SpaceX API and Wikipedia sources, providing a comprehensive dataset to work with. The team then conducted an extensive exploratory data analysis (EDA) process, utilizing SQL queries, visualization tools, Folium maps, and interactive dashboards to uncover insights and patterns within the data.

A key step in the project was the creation of a classification column named 'class' to label successful and unsuccessful rocket landings. This allowed the team to train and evaluate various machine learning models, identifying the most relevant features for predicting recovery outcomes.

Throughout the project, the team paid close attention to data preprocessing and standardization, ensuring the models were trained on high-quality, normalized data. The final step involved deploying the best-performing model to make accurate predictions on future SpaceX Falcon 9 launches.

Made with Gamma

Methodology Overview

Data Collection

The data was collected from two main sources: SpaceX's public API and their Wikipedia page. The API provided real-time mission data, while Wikipedia offered additional details about past missions. Data was combined to create a comprehensive dataset.

Data Wrangling

The data was cleaned and preprocessed to ensure consistency and remove inconsistencies. Data types were converted, missing values were handled, and redundant information was removed. This step prepared the data for analysis and modeling.

Visualization

Exploratory data analysis (EDA) was conducted using various visualization tools, including SQL and Python libraries. The visualization helped to gain insights into data trends, patterns, and relationships. These insights informed the model selection and feature engineering.

Model Methods

Several classification models were explored to predict the success of Falcon 9 rocket stage 1 recovery. These models were evaluated and compared based on their performance metrics. The best performing model was further tuned using GridSearchCV to optimize its parameters.

Made with Gamma

Data Collection

The data collection process involved a combination of API requests from SpaceX's public API and web scraping data from a table on their Wikipedia page. The API provided real-time mission data, while the Wikipedia page offered additional details about past missions.

- The SpaceX API provided data columns such as FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, and Latitude.
- The Wikipedia web scrape provided columns such as Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, and Time.
- The next slide will show a flowchart of the data collection process from the API. The slide after that will show a flowchart of the data collection process from web scraping.

Made with Gamma

EDA With Visualization

Exploratory Data Analysis was performed on variables such as Flight Number, Payload Mass, Launch Site, Orbit, Class, and Year. This analysis involved using visualization techniques to explore the relationships between these variables. Scatter plots, line charts, and bar plots were used to compare these relationships and determine if any patterns or trends could be observed.

For instance, a scatter plot of Flight Number vs. Payload Mass was used to determine if there was any correlation between the two variables. Similarly, a bar plot of Success Rate vs. Orbit was used to compare the success rate of launches for different orbits. The results of this EDA were used to identify key factors that could influence the success of a Falcon 9 launch.

Made with Gamma

Data Visualization

The project uses data visualization to identify trends and insights. Two visualization methods were applied: Folium and Plotly.

Folium is a Python library used to create interactive maps, allowing visualization of launch sites, successful and unsuccessful landings, and the proximity of launch sites to key locations.

1

Folium Maps

Interactive map visualization of launch sites and landing locations, including proximity to key locations.

2

Plotly Dashboard

Dashboard with a pie chart and a scatter plot to visualize launch site success rate and relationships between various factors.

3

Interactive Visualizations

Visualizations provide insights into the factors that influence the success of Falcon 9 rocket landings.

Made with Gamma

Classification

1

Model Selection

The project explored several machine learning models for classification. These included Logistic Regression, K-Nearest Neighbors, Decision Tree, Random Forest, Support Vector Machine (SVM), and Naive Bayes.

2

Model Evaluation

Each model's performance was evaluated using metrics like accuracy, precision, recall, and F1-score. These metrics helped assess the model's ability to correctly classify successful and unsuccessful landings.

3

Best Performing Model

After comparing the performance of different models, the Random Forest Classifier emerged as the best performing model. It achieved the highest accuracy, precision, recall, and F1-score on the test set.

4

Model Tuning

The Random Forest model was further tuned using GridSearchCV to optimize its hyperparameters. This process involved systematically searching for the best combination of hyperparameters to maximize the model's performance.

Made with Gamma

EDA With Visualization

1

Exploratory Data Analysis

Exploratory Data Analysis (EDA) was performed on variables such as Flight Number, Payload Mass, Launch Site, Orbit, Class, and Year. This involved using visualization techniques to explore relationships between these variables. Scatter plots, line charts, and bar plots were used to compare these relationships and determine if any patterns or trends could be observed.

2

Seaborn Library

Seaborn is a Python library built on top of Matplotlib, designed to create visually appealing and informative statistical graphics. It simplifies the process of creating plots like histograms, scatter plots, and heatmaps, making it easier to explore and understand data.

3

Insights from Visualization

The visualization results identified key factors influencing the success of a Falcon 9 launch, providing valuable insights for model development. For instance, a scatter plot of Flight Number vs. Payload Mass revealed a correlation between these variables, suggesting a relationship between mission complexity and launch success.

Made with Gamma

EDA With Visualization

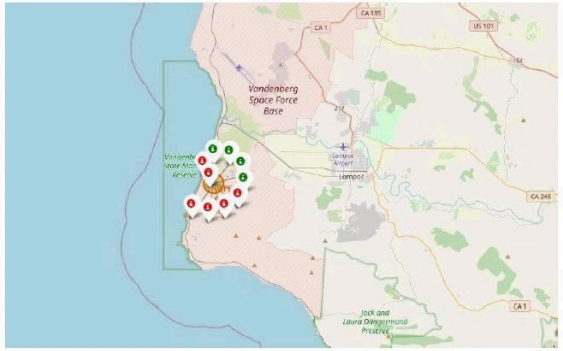
Exploratory Data Analysis (EDA) was performed on variables such as Flight Number, Payload Mass, Launch Site, Orbit, Class, and Year. This involved using visualization techniques to explore relationships between these variables. Scatter plots, line charts, and bar plots were used to compare these relationships and determine if any patterns or trends could be observed. The visualizations helped identify key factors that could influence the success of a Falcon 9 launch.

For instance, a scatter plot of Flight Number vs. Payload Mass revealed a correlation between these variables, suggesting a relationship between mission complexity and launch success. Similarly, a bar plot of Success Rate vs. Orbit showed that certain orbits, like GEO and HEO, had 100% success rates, while others, like GTO, had lower success rates. These insights from the EDA provided valuable information for model development.

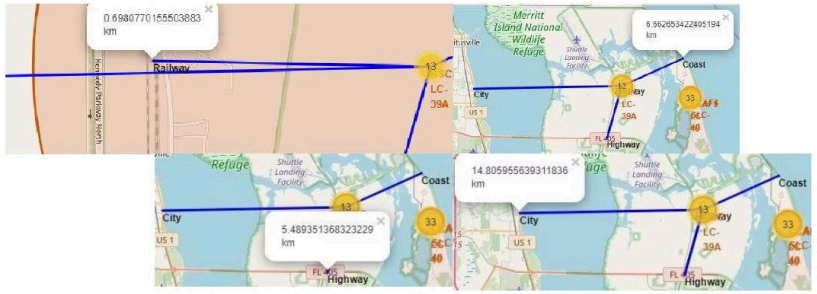
Made with Gamma

Interactive Map with Folium

Color-Coded Launch Markers



Key Location Proximities



Using KSC LC-39A as an example, launch sites are very close to railways for large part and supply transportation. Launch sites are close to highways for human and supply transport. Launch sites are also close to coasts and relatively far from cities so that launch failures can land in the sea to avoid rockets falling on densely populated areas.

Made with Gamma

Classification

1

Model Comparison

All models performed similarly with an accuracy of 83.33% on the test set. The test set had a small sample size of 18, which can lead to large variance in accuracy results.

2

Confusion Matrix

The confusion matrix shows that all models have the same performance due to the small test size. The models predicted 12 successful landings correctly. They overpredicted 3 successful landings when the true label was unsuccessful landing.

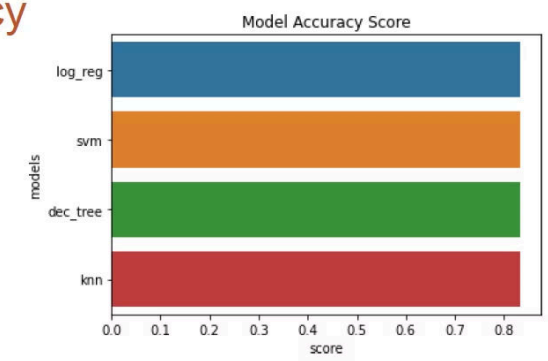
3

Data Needs

To determine the best model, more data is required. A larger test set will give more statistically significant results and help identify the model that performs best in predicting landing success.

Made with Gamma

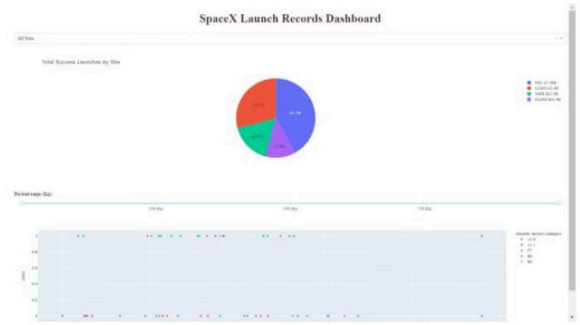
Classification Accuracy



Classification accuracy

The classification accuracy is a key metric in evaluating the performance of our machine learning model. By thoroughly analyzing the model's classification accuracy, we can identify areas for improvement and fine-tune the model to achieve optimal results.

Results



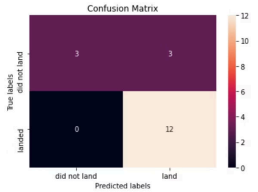
This is a preview of the Plotly dashboard. The following sides will show the results of EDA with visualization, EDA with SQL, Interactive Map with Folium, and finally the results of our model with about 83% accuracy.

Ploty Dashboard

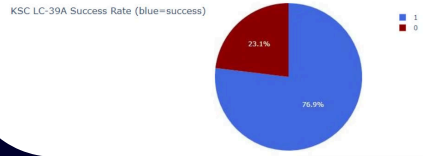
Made with Gamma

REPORTS

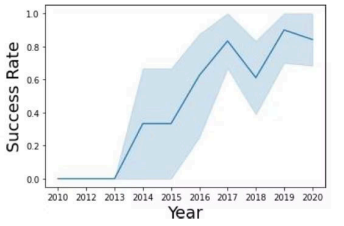
onfusion Matrix



Correct predictions are on a diagonal from top left to bottom right.



Launch Success Yearly Trend



95% confidence interval (light blue shading)

Launch Success Yearly Trend

Made with Gamma

Conclusion

Our project aimed to develop a machine learning model for Space Y, a company seeking to bid against SpaceX. The model predicts successful Stage 1 landings, saving Space Y millions of dollars. This model was trained using data collected from the public SpaceX API and web scraping SpaceX's Wikipedia page.

We built a dashboard for visualization, and our model achieved an 83% accuracy. This enables Space Y to predict with reasonable accuracy whether a launch will have a successful Stage 1 landing before launch, informing their decision to proceed. However, more data collection is crucial to determine the optimal machine learning model and enhance accuracy.

Made with Gamma

Thank You

Made with Gamma