IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

HAFEZ ELFAKIE

10 May 2024

Outline

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

Executive Summary

This study aims to uncover the key factors contributing to a successful rocket landing. To achieve this objective, the research employs the following methodologies:

Collect data using SpaceX REST API and web scraping techniques

- Wrangle data to create success/fail outcome variable

- **Explore data with data visualization techniques**, considering the following factors:

payload, launch site, flight number and yearly trend

- **Analyze the data with SQL**, calculating the following statistics: total payload, payload range for successful launches, and total # of successful and failed outcomes

Executive Summary (Cont.)

- **Explore** launch site success rates and proximity to geographical markers
- **Visualize** the launch sites with the most success and successful payload ranges
- **Build Models** to predict landing outcomes using logistic regression, support vector machine (SVM), decision tree and K-nearest neighbor (KNN)

Results

Exploratory Data Analysis

- **Launch success has improved over time**
- **KSC LC-39A has the highest success rate among landing sites**
- **Orbits ES-L1, GEO, HEO, and SSO have a 100% success rate**

Executive Summary (Results Cont.)

Visualization/Analytics

- **Most launch sites are near the equator, and all are close to the coast**

Predictive Analytics

- **All models performed similarly on the test set. The decision tree model slightly outperformed**

Introduction

Background

SpaceX, a frontrunner in space exploration, aims to bring down the cost of spaceflight. Their achievements include delivering spacecraft to the International Space Station, deploying an internet-beaming satellite network, and conducting human space missions. This affordability stems from their innovative approach to rocket reusability. SpaceX's Falcon 9 rocket boasts a reusable first stage, significantly reducing launch costs (around $62 million) compared to traditional rockets ($165 million and upwards) that can't be recovered. Predicting the reusability of the first stage, and consequently the launch price, is possible by analyzing publicly available data. Machine learning models can be trained to assess the likelihood of SpaceX – or any competitor – successfully recovering the first stage of their launch vehicle.

Introduction (Explore)

- How payload mass, launch site, number of flights, and orbits affect first-stage landing success
- Rate of successful landings over time
- Best predictive model for successful landing (binary classification)

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology
  - Using SpaceX REST API and web scraping techniques
- Perform data wrangling
  - By filtering the data, handling missing values - prepare the data for analysis and modeling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
- To predict landing outcomes using classification models. Tune and evaluate models to find best model and parameters

Data Collection

Data Collection – SpaceX API

- **Request data** from SpaceX API (rocket launch data)

- **Decode** response using .json() and convert to a dataframe using .json_normalize()

- **Request** information about the launches from SpaceX API using custom functions

- **Create dictionary** from the data

- **Create dataframe** from the dictionary

- **Filter dataframe** to contain only Falcon 9 launches

-**Replace missing values** of Payload Mass with calculated .mean()

- **Export data to csv file**

[Data Collection GitHub Link]

Data Collection - WebScraping

Request data (Falcon 9 launch data) from Wikipedia

- **Create BeautifulSoup object** from HTML response
- **Extract** column names from HTML table header
- **Collect data** from parsing HTML tables
- **Create dictionary** from the data
- **Create dataframe** from the dictionary
- **Export** data to csv file

Data Collection-WebScraping Github Link

Data Wrangling

- Perform EDA and determine data labels
- Calculate:
- • # of launches for each site
- • # and occurrence of orbit
- • # and occurrence of mission
- outcome per orbit type]
- Create binary landing outcome
- column (dependent variable)
- Export data to csv file

Data Wrangling (Landing Outcome Cont.)

Landing Outcome

- Landing was not always successful
- **True Ocean:** mission outcome had a successful landing to a specific region of the ocean 2023
- **False Ocean:** represented an unsuccessful landing to a specific region of ocean
- **True RTLS**: meant the mission had a successful landing on a ground pad
- **False RTLS:** represented an unsuccessful landing on a ground pad
- **True ASDS**: meant the mission outcome had a successful landing on a drone ship
- **False ASDS**: represented an unsuccessful landing on drone ship
- **Outcomes converted into 1 for a successful landing and 0 for an unsuccessful landing**

- [Data Wrangling GitHub Link]

EDA with Data Visualization

Charts

- Flight Number vs. Payload
- Flight Number vs. Launch Site
- Payload Mass (kg) vs. Launch Site
- Payload Mass (kg) vs. Orbit type

Analysis

- View relationship by using scatter plots. The variables could be useful for machine learning if a relationship exists
- Show comparisons among discrete categories with bar charts. Bar charts show the relationships among the categories and a measured value.

EDA With Data Visualization GitHub Link

EDA with SQL

Queries

Display:

- Names of unique launch sites
- 5 records where launch site begins with 'CCA'
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1.

List:

- Date of first successful landing on ground pad
- Names of boosters which had success landing on drone ship and have payload mass greater than 4,000 but less than 6,000
- Total number of successful and failed missions
- Names of booster versions which have carried the max payload
- Failed landing outcomes on drone ship, their booster version and launch site for the months in the year 2015
- Count of landing outcomes between 2010-06-04 and 2017-03-20 (desc)

EDA With SQL GitHub Link

Build an Interactive Map with Folium

Markers Indicating Launch Sites

- • Added blue circle at NASA Johnson Space Center's coordinate with a popup label showing its name using its latitude and longitude coordinates
- • Added red circles at all launch sites coordinates with a popup label showing its name using its name using its latitude and longitude coordinates
- Colored Markers of Launch Outcomes
- • Added colored markers of successful (green) and unsuccessful (red)
- launches at each launch site to show which launch sites have high success
- rates
- Distances Between a Launch Site to Proximities
- • Added colored lines to show distance between launch site CCAFS SLC- 40 and its proximity to the nearest coastline, railway, highway, and city
- Folium GitHub Link

Build a Dashboard with Plotly Dash

Dropdown List with Launch Sites

- Allow user to select all launch sites or a certain launch site

Pie Chart Showing Successful Launches

- Allow user to see successful and unsuccessful launches as a percent of the total

Slider of Payload Mass Range

- Allow user to select payload mass range

Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version

- Allow user to see the correlation between Payload and Launch Success

DashBoard GitHub Link

Predictive Analysis (Classification)

- Charts
- Create NumPy array from the Class column
- Standardize the data with StandardScaler. Fit and transform the data.
- Split the data using train_test_split
- Create a GridSearchCV object with cv=10 for parameter optimization
- Apply GridSearchCV on different algorithms: logistic regression (LogisticRegression()), support vector machine (SVC()), decision tree (DecisionTreeClassifier()), K-Nearest Neighbor (KNeighborsClassifier())
- Calculate accuracy on the test data using .score() for all models
- Assess the confusion matrix for all models
- Identify the best model using Jaccard_Score, F1_Score and Accuracy
- Machine Learning GitHub Link

Results

Exploratory Data Analysis

- Launch success has improved over time
- KSC LC-39A has the highest success rate among landing sites
- Orbits ES-L1, GEO, HEO and SSO have a 100% success rate

Visual Analytics

- Most launch sites are near the equator, and all are close to the coast
- Launch sites are far enough away from anything a failed launch can damage (city, highway, railway), while still close enough to bring people and material to support launch activities

Predictive Analytics

- Decision Tree model is the best predictive model for the dataset

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

Exploratory Data Analysis

- Earlier flights had a lower success rate (blue = fail)
- Later flights had a higher success rate (orange = success)
- Around half of launches were from CCAFS SLC 40 launch site
- VAFB SLC 4E and KSC LC 39A have higher success rates
- We can infer that new launches have a higher success rate

Payload vs. Launch Site

- **Exploratory Data Analysis**
- **Typically, the higher the payload mass (kg), the higher the success rate**
- **Most launces with a payload greater than 7,000 kg were successful**
- **KSC LC 39A has a 100% success rate for launches less than 5,500 kg**
- **VAFB SKC 4E has not launched anything greater than ~10,000 kg**

Success Rate vs. Orbit Type

Exploratory Data Analysis

- 100% Success Rate: ES-L1, GEO, HEO and SSO
- 50%-80% Success Rate: GTO, ISS, LEO, MEO, PO
- 0% Success Rate: SO

Flight Number vs. Orbit Type

- **Exploratory Data Analysis**
- **The success rate typically increases with the number of flights for each orbit**
- **This relationship is highly apparent for the LEO orbit**
- **The GTO orbit, however, does not follow this trend**

Payload vs. Orbit Type

- **Exploratory Data Analysis**
- **Heavy payloads are better with LEO, ISS and PO orbits**
- **The GTO orbit has mixed success with heavier payloads**

Launch Success Yearly Trend

- **Exploratory Data Analysis**
- **The success rate improved from 2013-2017 and 2018-2019**
- **The success rate decreased from 2017-2018 and from 2019-2020**
- **Overall, the success rate has improved since 2013**

All Launch Site Names

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

In [11]:

```
t1 = pd.read_sql('select distinct Launch_Site from SPACEXTBL', con)
t1
```

Out[11]:

	Launch_Site
0	CCAFS LC-40
1	VAFB SLC-4E
2	KSC LC-39A
3	CCAFS SLC-40

Launch Site Names Begin with 'CCA'

In [15]:
```
t2 = pd.read_sql("select * from SPACEXTBL where Launch_Site like 'CCA%' limit 5", con)
t2
```

Out[15]:

	Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
0	2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
1	2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of...	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2	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
3	2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
4	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

Total Payload MassAverage Payload Mass

- **45,596 kg (total) carried by boosters launched by NASA (CRS)**

In [16]:

```
t3 = pd.read_sql("select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer='NASA (CRS)'", con)
t3
```

Out[16]:

	sum(PAYLOAD_MASS__KG_)
0	45596

Average Payload Mass by F9 v1.1

Average Payload Mass

2,928 kg (average) carried by booster version F9 v1.1

In [17]:
```
t4 = pd.read_sql("select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version='F9 v1.1'", con)
t4
```

Out[17]:

	avg(PAYLOAD_MASS__KG_)
0	2928.4

First Successful Ground Landing Date

1st Successful Landing in Ground Pad

- **12/22/2015**

In [19]:
```
t5 = pd.read_sql("select min(Date) from SPACEXTBL where Landing_Outcome='Success (ground pad)'", con)
t5
```

Out[19]:

	min(Date)
0	2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Booster Drone Ship Landing

- Booster mass greater than 4,000 but less than 6,000
- F9 FT B1022 , F9 FT B1026, F9 FT B1021.2, F9 FT B1031.2

In [23]:

```
t6= pd.read_sql("select distinct Booster_Version from SPACEXTBL \
                 where Landing_Outcome='Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 and 6000", con
t6
```

Out[23]:

	Booster_Version
0	F9 FT B1022
1	F9 FT B1026
2	F9 FT B1021.2
3	F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

Total Number of Successful and Failed Mission Outcomes

- 1 Failure in Flight
- 100 Success

In [24]:

```
t7 = pd.read_sql("select substr(Mission_Outcome,1,7) as Mission_Outcome, count(*) \
from SPACEXTBL  group by 1", con)
t7
```

Out[24]:

	Mission_Outcome	count(*)
0	Failure	1
1	Success	100

Boosters Carried Maximum Payload

In [25]:

```
t8 = pd.read_sql("select distinct Booster_Version from SPACEXTBL \
where PAYLOAD_MASS__KG_= (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)", con)
t8
```

Out[25]:

	Booster_Version
0	F9 B5 B1048.4
1	F9 B5 B1049.4
2	F9 B5 B1051.3
3	F9 B5 B1056.4
4	F9 B5 B1048.5
5	F9 B5 B1051.4
6	F9 B5 B1049.5
7	F9 B5 B1060.2
8	F9 B5 B1058.3
9	F9 B5 B1051.6
10	F9 B5 B1060.3
11	F9 B5 B1049.7

2015 Launch Records

In [27]:

```
t9 = pd.read_sql("select distinct Landing_Outcome, Booster_Version,Launch_Site,substr(Date,0,5)='2015'  \
from SPACEXTBL where Landing_Outcome='Failure (drone ship)'", con)
t9
```

Out[27]:

	Landing_Outcome	Booster_Version	Launch_Site	substr(Date,0,5)='2015'
0	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	1
1	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	1
2	Failure (drone ship)	F9 v1.1 B1017	VAFB SLC-4E	0
3	Failure (drone ship)	F9 FT B1020	CCAFS LC-40	0
4	Failure (drone ship)	F9 FT B1024	CCAFS LC-40	0

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

In [29]:

```
t10 = pd.read_sql("select Landing_Outcome, count(*) from SPACEXTBL \
where Date between '2011-06-04' and '2017-03-20' group by Landing_Outcome order by 2 desc", con)
t10
```

Out[29]:

	Landing_Outcome	count(*)
0	No attempt	10
1	Success (drone ship)	5
2	Failure (drone ship)	5
3	Success (ground pad)	3
4	Controlled (ocean)	3
5	Uncontrolled (ocean)	2
6	Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

Launch Sites

- **Near Equator**: the closer the launch site to the equator, the **easier it is to launch** to equatorial orbit, and the more help you get from Earth's rotation for a prograde orbit. Rockets launched from sites near the equator get an additional natural boost - due to the rotational speed of earth - that helps save the cost of putting in extra fuel and boosters.

Launch Outcomes

- At Each Launch Site
- • Outcomes:
- Green markers for successful launches
- Red markers for unsuccessful launches
- Launch site CCAFS SLC-40 has a 3/7 success rate (42.9%)

Distance to Proximities

- **CCAFS SLC-40**
- **• .86 km from nearest coastline**
- **• 21.96 km from nearest railway**
- **• 23.23 km from nearest city**
- **• 26.88 km from nearest highway**

Section 4

Build a Dashboard with Plotly Dash

Launch Success by Site Success as Percent of Total

Success as Percent of Total

- **KSC LC-39A has the most successful launches amongst launch sites (41.2%)**

Launch Success (KSC LC-29A) Success as Percent of Total

- **KSC LC-39A has the highest success rate amongst launch sites (76.9%)**
- **10 successful launches and 3 failed launches**

Payload Mass and Success By Booster Version

- **Payloads between 2,000 kg and 5,000 kg have the highest success rate**
- **1 indicating successful outcome and 0 indicating an unsuccessful outcome**

Section 5

Predictive Analysis (Classification)

Classification Accuracy

All the models performed at about the same level and had the same scores and accuracy. This is likely due to the small dataset. The Decision Tree model slightly outperformed the rest when looking at .best_score_

- **.best_score_ is the average of all cv folds for a single combination of the parameters**

```python
models = {'KNeighbors':knn_cv.best_score_,
              'DecisionTree':tree_cv.best_score_,
              'LogisticRegression':logreg_cv.best_score_,
              'SupportVector': svm_cv.best_score_}

bestalgorithm = max(models, key=models.get)
print('Best model is', bestalgorithm,'with a score of', models[bestalgorithm])
if bestalgorithm == 'DecisionTree':
    print('Best params is :', tree_cv.best_params_)
if bestalgorithm == 'KNeighbors':
    print('Best params is :', knn_cv.best_params_)
if bestalgorithm == 'LogisticRegression':
    print('Best params is :', logreg_cv.best_params_)
if bestalgorithm == 'SupportVector':
    print('Best params is :', svm_cv.best_params_)
```

```
Best model is DecisionTree with a score of 0.8892857142857145
Best params is : {'criterion': 'entropy', 'max_depth': 8, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samp
les_split': 10, 'splitter': 'random'}
```

Confusion Matrix

- Performance Summary
- A confusion matrix summarizes the performance of a classification algorithm
  - All the confusion matrices were identical
- The fact that there are false positives (Type 1 error) is not good
- Confusion Matrix Outputs:
- 12 True positive
- 3 True negative
- 3 False positive
- 0 False Negative
- Precision = TP / (TP + FP)
- 12 / 15 = .80
- Recall = TP / (TP + FN)
- 12 / 12 = 1
- F1 Score = 2 * (Precision * Recall) / (Precision + Recall)
- 2 * (.8 * 1) / (.8 + 1) = .89
- Accuracy = (TP + TN) / (TP + TN + FP + FN) = .833

Conclusions

Model Performance: The models performed similarly on the test set with the decision tree model slightly outperforming

- **Equator**: Most of the launch sites are near the equator for an additional natural boost - due to the rotational speed of earth - which helps save the cost of putting in extra fuel and boosters
- **Coast:** All the launch sites are close to the coast
- **Launch Success**: Increases over time
- **KSC LC-39A**: Has the highest success rate among launch sites. Has a 100% success rate for launches less than 5,500 kg
- **Orbits**: ES-L1, GEO, HEO, and SSO have a 100% success rate
- **Payload Mass**: Across all launch sites, the higher the payload mass (kg), the higher the success rate

Thank you!