



APPLIED DATA SCIENCE CAPSTONE

PEIWEN GUAN

NOVEMBER 24, 2022

GITHUB LINK: [COURSERA/APPLIED DATA SCIENCE CAPSTONE AT MAIN • HI2E3G/COURSERA \(GITHUB.COM\)](https://github.com/HI2E3G/COURSERA-APPLIED-DATA-SCIENCE-CAPSTONE-AT-MAIN)

TABLE OF CONTENTS

- Project Introduction
- Data Collection and Data Wrangling
- Exploratory Data Analysis (EDA)
- Interactive Visual Analytics and Dashboard
- Predictive Analysis (Classification)
- Conclusion
- Reference





APPLIED DATA SCIENCE CAPSTONE

BY IBM SKILLS NETWORK

This is the final course in the IBM Data Science Professional Certificate as well as the Applied Data Science with Python Specialization. Student will practice the work that data scientists do in real life when working with datasets.

In this course student will assume the role of a Data Scientist working for a startup intending to compete with SpaceX, and in the process follow the Data Science methodology involving data collection, data wrangling, exploratory data analysis, data visualization, model development, model evaluation, and reporting your results to stakeholders.



DATA COLLECTION AND DATA WRANGLING

- Public API: API requests from Space X
 - Web Scraping: Wikipedia page about Space X
 - Data Wrangling: Land outcomes to be converted to Classes y (either 0 or 1). 0 is a bad outcome, that is, the booster did not land. 1 is a good outcome, that is, the booster did land.
- Public API Methodology: Request (API) → Json → Data Frame → Data filtering (Falcon 9) → Data cleaning (using mean value for missing Payload Mass values)
 - Web scraping Methodology: Request (html) → HTML parse → Extract key values → Dictionary format



EXPLORATORY DATA ANALYSIS (EDA)

- **Collect data on the Falcon 9 first-stage landings**
- **Create scatter plots and bar charts by writing Python code to analyze data in a Pandas data frame**
- **Write Python code to conduct exploratory data analysis by manipulating data in a Pandas data frame**
- **Write and execute SQL queries to select and sort data**
- **Use your data visualization skills to visualize the data and extract meaningful patterns to guide the modeling process.**

EXPLORATORY DATA ANALYSIS (EDA) RESULTS

Display the names of the unique launch sites in the space

```
%sql select DISTINCT LAUNCH_SITE from SPACEXTBL
```

* ibm_db_sa://mmp08973:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/BLUDB
Done.

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

Display a record where launch sites begin with the string CCA

```
%sql select * from SPACEXTBL where launch_site like 'CCA%' limit 5
```

* ibm_db_sa://mmp08973:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/BLUDB
Done.

DATE	Time (UTC)	booster_version	launch_site	payload	payload_mass_kg	orbit	customer	mission_outcome	Landing Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachut
2010-08-12	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 (parachut
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No atten
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No atten
2013-03-12	22:41:00	F9 v1.1	CCAFS LC-40	SES-8	3170	GTO	SES	Success	No atten

```
%sql select sum(payload_mass_kg_) as sum from SPACEXTBL where customer like 'NASA (CRS)'
```

* ibm_db_sa://mmp08973:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/BLUDB
Done.

SUM
22007

Display average payload mass carried by booster version F9 v1.1

```
%sql select avg(payload_mass_kg_) as Average from SPACEXTBL where booster_version like 'F9 v1.1'
```

* ibm_db_sa://mmp08973:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/BLUDB
Done.

average
3226

```
%sql  
SELECT booster_version  
FROM SPACEXTBL  
WHERE landing_outcome = 'Success (drone ship)' AND payload
```

* ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b52
Done.

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

```
maxm = %sql select max(payload_mass_kg_) from SPACEXTBL  
maxv = maxm[0][0]
```

```
%sql select booster_version from SPACEXTBL where payload_mass_kg_=(select max(payload_mass_kg_) from SPACEXTBL)
```

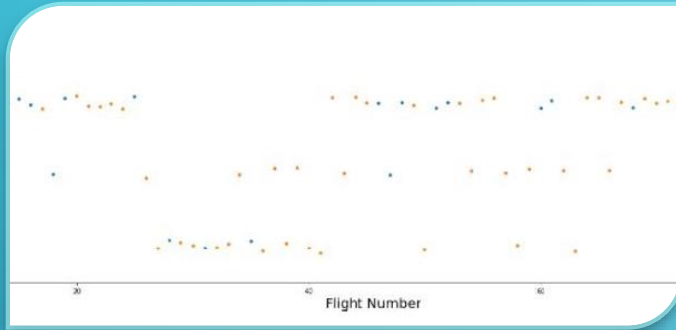
* ibm_db_sa://mmp08973:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/BLUDB
Done.
* ibm_db_sa://mmp08973:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/BLUDB
Done.

booster_version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3

INTERACTIVE VISUAL ANALYTICS AND DASHBOARD

- Build an interactive dashboard that contains pie charts and scatter plots to analyze data with the Plotly Dash Python library
- Calculate distances on an interactive map by writing Python code using the Folium library
- Generate interactive maps, plot coordinates, and mark clusters by writing Python code using the Folium library
- Build a dashboard to analyze launch records interactively with Plotly Dash.
- Build an interactive map to analyze the launch site proximity with Folium.

INTERACTIVE VISUAL ANALYTICS AND DASHBOARD



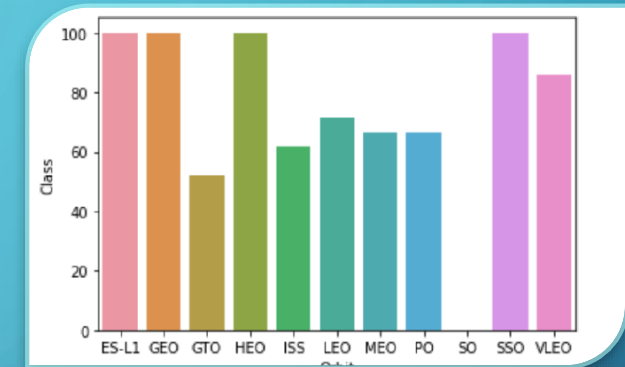
FLIGHT NUMBER VS. LAUNCH SITE

Orange indicates successful launch, and blue indicates unsuccessful launch. CCAFS appears to be the main launch site as it has the most volume.



PAYLOAD VS. LAUNCH SITE

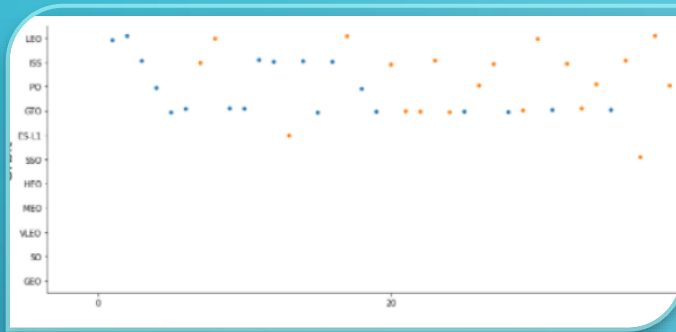
Orange indicates successful launch, and blue indicates unsuccessful launch. Payload mass appears to fall mostly between 0-7000kg. Different launch sites seems to use different payload mass.



SUCCESS RATE VS. ORBIT TYPE

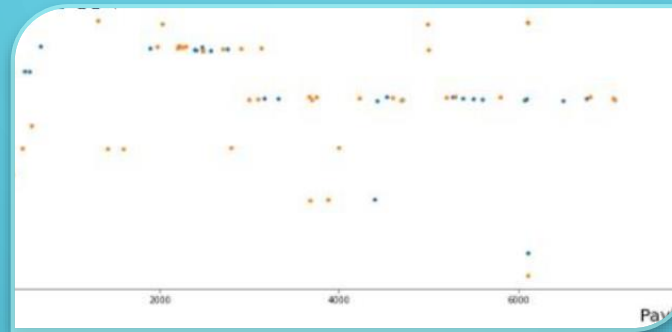
Based on the graphic, we can know ES-L1(1), GEO(1), HEO(1) have 100% success rate.

INTERACTIVE VISUAL ANALYTICS AND DASHBOARD



FLIGHT NUMBER VS. ORBIT TYPE

Launch Outcome seems to correlate with this preference. SpaceX appears to perform better in lower orbits or Sun-synchronous orbits.



PAYLOAD VS. LAUNCH SITE

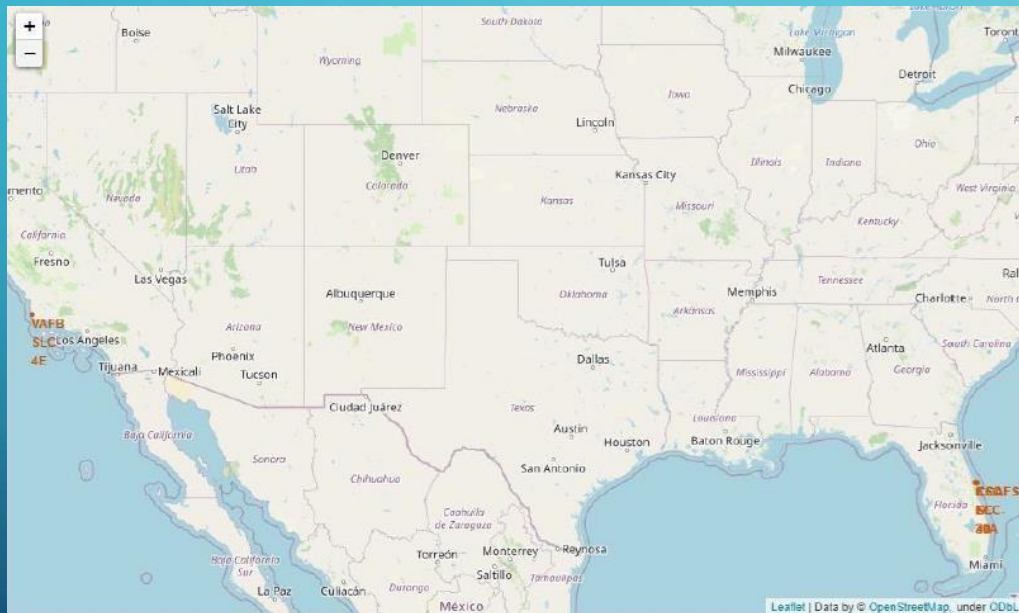
Payload mass seems to correlate with orbit. LEO and SSO seem to have relatively low payload mass



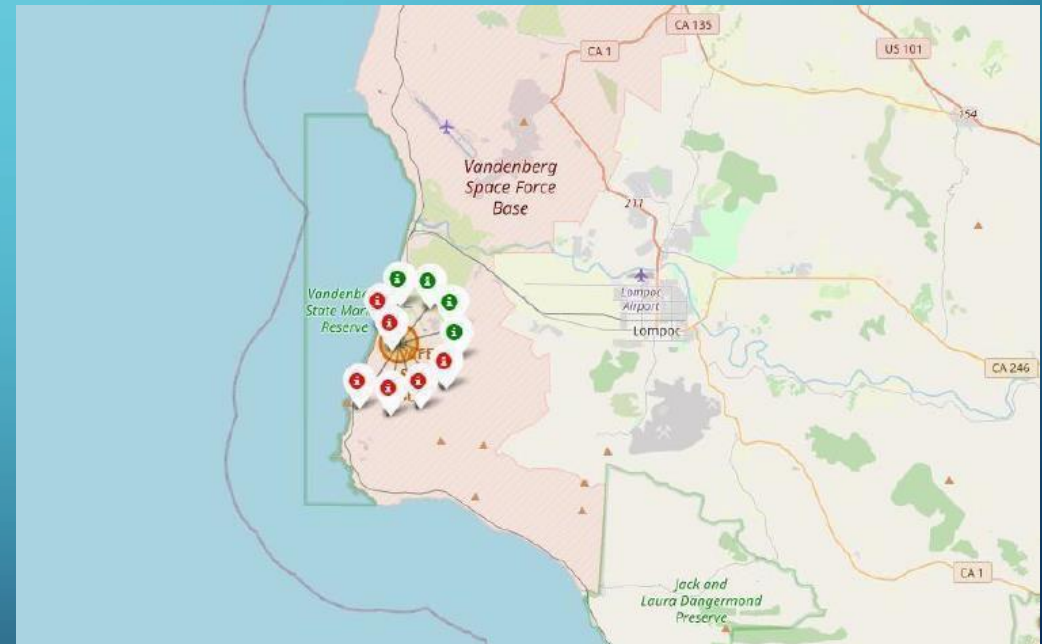
LAUNCH SUCCESS YEARLY TREND

After 2013, launch success generally increasing, which rate is almost 85%.

INTERACTIVE MAP WITH FOLIUM

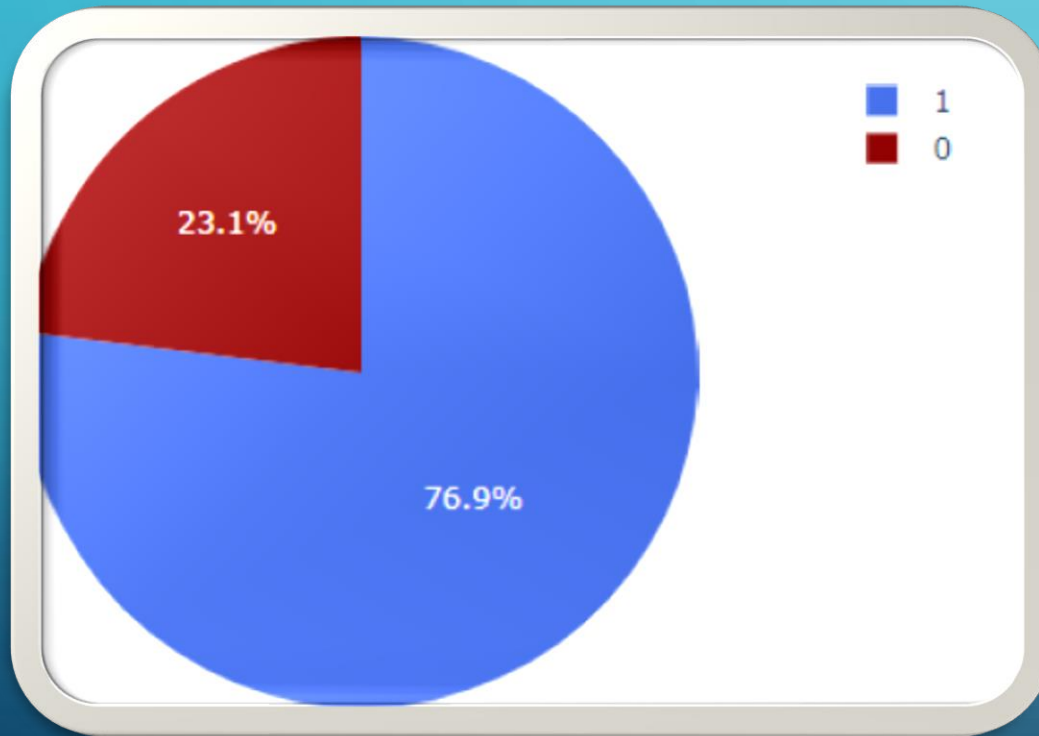


Launch Site Locations in the U.S.



VAFBSLC-4E shows 4 times successful landings and 6 failed landings.

INTERACTIVE VISUAL ANALYTICS AND DASHBOARD



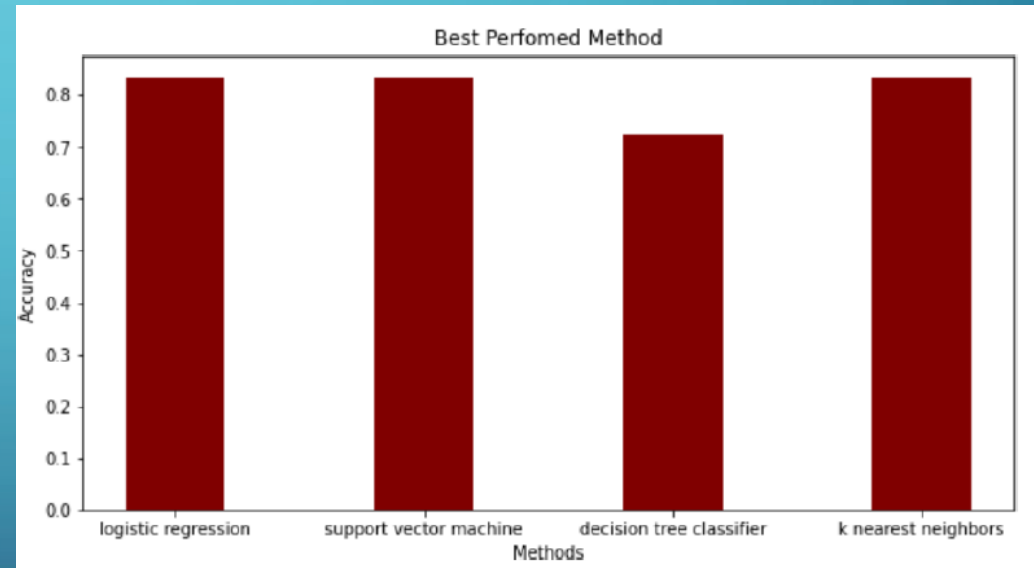
From the left pie chart, we can see KSC LC-39A has the highest success rate with 10 successful landings and 3 failed landings.

PREDICTIVE ANALYSIS (CLASSIFICATION)

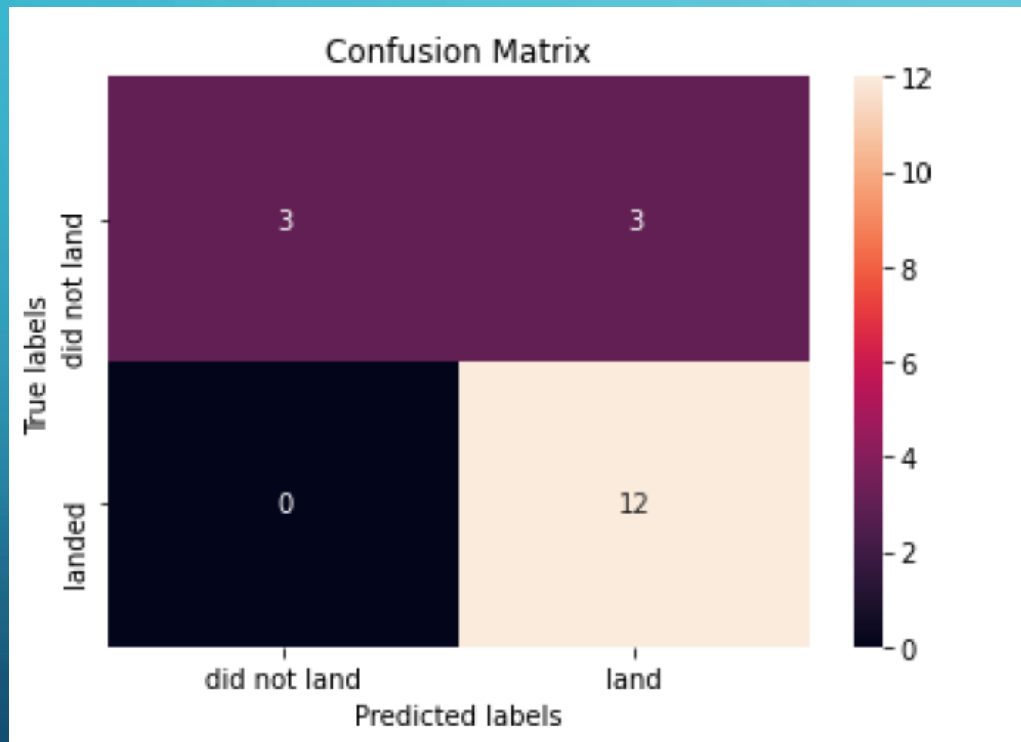
- Split the data into training testing data.
- Train different classification models.
- Hyperparameter grid search.
- Use your machine learning skills to build a predictive model to help a business function more efficiently.

PREDICTIVE ANALYSIS (CLASSIFICATION)

- The models had virtually the same accuracy on the test set at 83.33% accuracy, except the decision tree classifier with 72,23%.
- For more accurate prediction, we need more data for analyzing and train our models.



PREDICTIVE ANALYSIS (CLASSIFICATION)



- A confusion matrix is a tabular summary of the number of correct and incorrect predictions made by a classifier.
- The model predicted 12 successful landings when the true label was successful landing.

CONCLUSION

- In this course I am trying to be a data scientist by using the data from SpaceX, and in the process follow the Data Science methodology involving data collection, data wrangling, exploratory data analysis, data visualization, model development, model evaluation.

REFERENCE

- [Coursera/Applied Data Science Capstone at main · hi2e3g/Coursera \(github.com\)](#)
- [Applied Data Science Capstone - Introduction | Coursera](#)