Data Science Capstone

Will Falcon 9 First Stage Rockets Land Successfully?

STEPHANIE ORGILL AUGUST 2021

Outline



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

Executive Summary

FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs
1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False
2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False
3	2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False
4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False
5	2013-12-03	Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	False	False
6	2014-01-06	Falcon 9	3325.000000	GTO	CCAFS SLC 40	None None	1	False	False	False
7	2014-04-18	Falcon 9	2296.000000	ISS	CCAFS SLC 40	True Ocean	1	False	False	True
8	2014-07-14	Falcon 9	1316.000000	LEO	CCAFS SLC 40	True Ocean	1	False	False	True
9	2014-08-05	Falcon 9	4535.000000	GTO	CCAFS SLC 40	None None	1	False	False	False
10	2014-09-07	Falcon 9	4428.000000	GTO	CCAFS SLC 40	None None	1	False	False	False

Methodology:

- Data collected from SpaceX API
- Data web scraped from SpaceX online tables that hold historical Falcon 9 launch information
- The data was collected from the two sources above and cleaned to hold only relevant information.
- Then the data was used for exploratory data analysis, visual analysis, and machine learning predictions to better understand if Falcon 9 rockets will land successfully.

Introduction



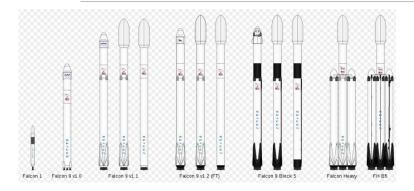
Background & context:

- SpaceX advertises Falcon 9 rocket launches at a cost of \$62 million, which is much less than the ~\$165 million of other providers.
- These savings largely come from the fact that SpaceX can reuse the first stage of rockets.
- Therefore, if we can predict if the first stage will land, we can also predict the cost of a rocket launch.
- This information would be useful to have if an alternate company wants to bid against SpaceX for a rocket launch.

Business Problem:

 Can we accurately predict if the first stage will land successfully?

Methodology



- Data collection methodology:
 - The data was collected via the SpaceX API and web scraped from online sources that hold historical launch success and failure data.
- Perform data wrangling
 - Once the data was acquired, it was cleansed using Python and SQL to:
 - remove irrelevant information
 - create dummy variables for categorical data
 - remove or replace missing information
- Perform exploratory data analysis (EDA) using visualization and SQL
 - Python and SQL were both used to:
 - · Identify launch sites
 - Find success and failure rates of Falcon 9 first launches
 - Identify relevant factors, such as booster versions, payload mass, and Orbit type
- Perform interactive visual analytics using Folium
 - Folium mapped the location of launch sites, NASA, and successes and failures at each location.
- Perform predictive analysis using classification models
 - Logistic Regression, Support Vector Machine, Decision Tree, and K-Nearest Neighbor models were all used (to roughly the same accuracy rate) to predict success rates of the Falcon 9 launches

Data collection

SpaceX API:

- •The SpaceX REST API was used to gather data regarding each launch.
- •The irrelevant information was removed from the dataset to keep only the information we wanted, such as the rocket type, payload, flight number, cores, etc.

Webscraping:

- •Using BeautifulSoup and Python, historical launch data was webscraped from online data sources.
- •The irrelevant information was removed to keep only the information required, such as flight number, launch site, payload mass, launch outcome, etc.

Data collection – SpaceX API

Add the GitHub URL of the completed SpaceX API calls notebook as an external reference and peer-review purpose.

Data Collection AP

create dataframe from launch_dict
data = pd.DataFrame(data = launch_dict)

show head of dataframe
data.head()

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	Reused Count	Serial	Longitude	Latitude
0	1	2006-03-24	Falcon 1	20.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin1A	167.743129	9.047721
1	2	2007-03-21	Falcon 1	NaN	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin2A	167.743129	9.047721
2	4	2008-09-28	Falcon 1	165.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin2C	167.743129	9.047721
3	5	2009-07-13	Falcon 1	200.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin3C	167.743129	9.047721
4	6	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857

Data collection – Web scraping

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

Data Collection Webscraping

df=	pd.DataFrame	({ key:pd.	Series(value) for key, value	in launch_di	ct.ite	ms() })					
df											
	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	flight_number	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Success\n	4 June 2010	18:45
1	flight_number	CCAFS	Dragon	0	LEO	NASA (COTS)\nNRO	Success	F9 v1.0B0004.1	Success	8 December 2010	15:43
2	flight_number	CCAFS	Dragon	525 kg	LEO	NASA (COTS)	Success	F9 v1.0B0005.1	Success	22 May 2012	07:44
3	flight_number	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA (CRS)	Success\n	F9 v1.0B0006.1	Success\n	8 October 2012	00:35
4	flight_number	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA (CRS)	Success\n	F9 v1.0B0007.1	Success\n	1 March 2013	15:10
	***	***	***								
116	flight_number	CCSFS	Starlink	15,600 kg	LEO	SpaceX	Success\n	F9 B5B1051.10	Success\n	9 May 2021	06:42
117	flight_number	KSC	Starlink	~14,000 kg	LEO	SpaceX Capella Space and Tyvak	Success\n	F9 B5B1058.8	Success\n	15 May 2021	22:56
118	flight_number	CCSFS	Starlink	15,600 kg	LEO	SpaceX	Success\n	F9 B5B1063.2	Success\n	26 May 2021	18:59
119	flight_number	KSC	SpaceX CRS-22	3,328 kg	LEO	NASA (CRS)	Success\n	F9 B5B1067.1	Success\n	3 June 2021	17:29
120	flight_number	CCSFS	SXM-8	7,000 kg	GTO	Sirius XM	Success\n	F9 B5	Success\n	6 June 2021	04:26

121 rows × 11 columns

Data wrangling

Using Python and SQL, the data from both the API and Webscraping collections was cleaned up in similar ways:

- The data was converted into dataframes and tables for easier processing
- Irrelevant information was removed from the dataframes and tables
- Missing information was either removed or replaced by the mean of the feature
- Categorical data was converted into useable numerical values via dummy variables and one-hot encoding

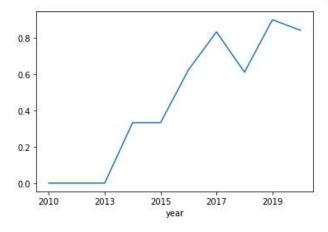
Data Wrangling

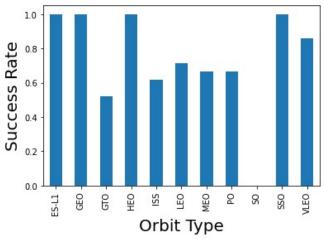
EDA with data visualization

EDA with visualization was used to:

- Explore the relationship between features, such as payload mass vs launch site
- Visualize the relationship between features, such as success rate and orbit type
- Visualize success rate over time

EDA with Visualization





EDA with SQL

SQL was used to identify helpful information such as:

- Launch sites
- Total & average payload mass (and related success / failure rates)
- Dates of successful landings
- The relationship between boosters and payload mass
- Total successful and failed launches : %%sql

EDA with SQL

```
%%sql

SELECT "Landing_Outcome", COUNT(*) AS TOTAL
FROM SPACEXTBL
WHERE (("Landing_Outcome" LIKE '%Success%') AND ("Date" > '2010-06-04') AND ("Date" < '2017-03-20'))
GROUP BY "Landing_Outcome"
ORDER BY "Landing_Outcome" DESC;

* ibm_db_sa://fgl32023:***@dashdb-txn-sbox-yp-dal09-04.services.dal.bluemix.net:50000/BLUDB
Done.
Landing_Outcome total
Success (ground pad) 3
Success (drone ship) 5</pre>
```

Build an interactive map with Folium

Folium was used to:

- Locate the launch sites and NASA on a map
- Identify successful and failed launches at each launch site
- Calculate distance from each launch site to its proximities

Interactive Visual Analytics





Predictive analysis (Classification)

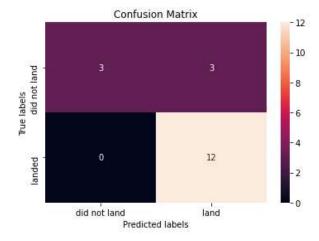
Machine Learning and Classification were used to predict the chances of success for Falcon 9 Launches.

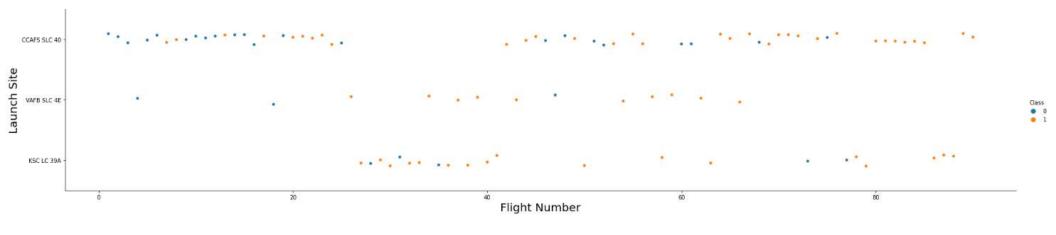
Four types of models were utilized, all with similar rates of accuracy (~84%):

- Logistic Regression
- Support Vector Machine
- Decision Tree
- K-Nearest Neighbors

These models were all pretty accurate, with their weakest points being false positives.

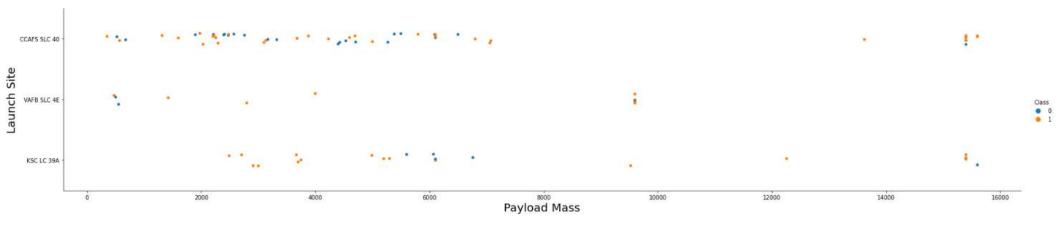
Predictive Analysis





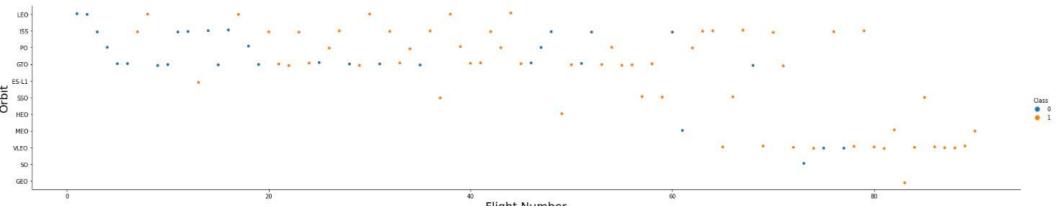
Flight Number vs. Launch Site

The higher the flight number, the more successful landing there are in every launch site.



Payload vs. Launch Site

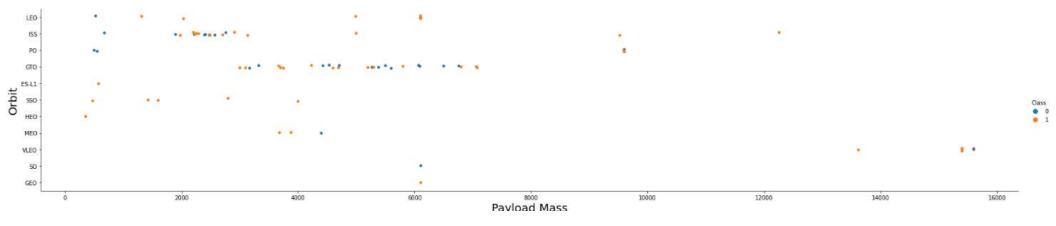
The CCAFS SLC-40 site has the most failures.



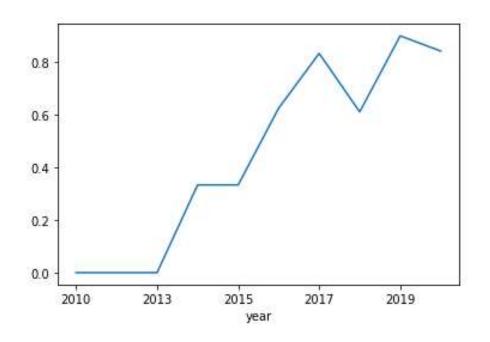
Flight Number vs. Orbit type

The higher the flight number, the more successes there tend to be.

The lower the orbit, the more successes there tends to be.



Payload vs. Orbit type



Launch success yearly trend

Launches have historically become more and more successful over time.

All launch site names

The four unique launch sites were found using SQL as below:

Task 1: Display the names of the unique launch sites in the space mission

```
%%sql
SELECT DISTINCT("Launch_Site")
FROM SPACEXTBL;

* ibm_db_sa://fgl32023:***@dashdb-txn-sbox-yp-dal09-04.services.dal.bluemix.np
Done.
Launch_Site
CCAFS LC-40
CCAFS SLC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E
```

Launch site names begin with `CCA`

Task 2: Display 5 records where launch sites begin with string 'CCA'

```
%%sql

SELECT *
FROM SPACEXTBL
WHERE "Launch_Site" LIKE 'CCA%'
LIMIT 5;

* ibm_db_sa://fgl32023:***@dashdb-txn-sbox-yp-dal09-04.services.dal.bluemix.net:50000/BLUDB
Done.
```

Date	Time_UTC_	Booster_Version	Launch_Site	Payload	payload_masskg_	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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

Total payload mass carried by NASA-launched boosters can be found below:

Task 3: Display the total payload mass carried by boosters launched by NASA (CRS)

Average payload mass by F9 v1.1

Average payload mass can be found below:

Task 4: Display average payload mass carried by booster version F9 v1.1

```
%%sql

SELECT AVG(payload_mass__kg_) AS AVG_MASS
FROM SPACEXTBL
WHERE "Booster_Version" LIKE 'F9 v1.1%';

* ibm_db_sa://fgl32023:***@dashdb-txn-sbox-yp-dal09-04.services.dal.bluemix.net:50000/BLUDB
Done.
    avg_mass
2534.666666
```

First successful ground landing date

The date of the first successful ground landing can be found below:

Task 5: List the date when the first successful landing outcome in ground pad was achieved

```
: %%sql

SELECT MIN("Date")
FROM SPACEXTBL
WHERE "Landing_Outcome" LIKE '%ground pad%';

* ibm_db_sa://fgl32023:***@dashdb-txn-sbox-yp-dal09-04.services.dal.bluemix.net:50000/BLUDB
Done.
: 1
2015-12-22
```

Successful drone ship landing with payload between 4000 and 6000

Successful drone ship landings between 4000 and 6000 can be found below:

Task 6: List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

Total number of successful and failure mission outcomes

Total successful and failed missions can be found below:

Task 7: List the total number of successful and failure mission outcomes

```
%%sql

SELECT COUNT(*),
SUM("Landing_Outcome" LIKE '%Success%') AS SUCCESSES,
SUM("Landing_Outcome" LIKE '%Failure%') AS FAILURES
FROM SPACEXTBL;

* ibm_db_sa://fgl32023:***@dashdb-txn-sbox-yp-dal09-04.services.dal.bluemix.r
Done.

1 successes failures

101 61 10
```

Boosters carried maximum payload

Boosters that carried max pavload mass can be found below:

Task 8: List the names of the booster_versions which have carried the maximum payload mass. L

```
SELECT "Booster_Version", payload_mass__kg_
FROM SPACEXTBL
WHERE payload_mass__kg_ = (SELECT MAX(payload_mass__kg_) FROM SPACEXTBL);
 * ibm_db_sa://fgl32023:***@dashdb-txn-sbox-yp-dal09-04.services.dal.bluemix.net:50000/BLUDB
Booster_Version payload_mass__kg_
  F9 B5 B1048.4
  F9 B5 B1049.4
                           15600
  F9 B5 B1051.3
                           15600
  F9 B5 B1056.4
                           15600
  F9 B5 B1048.5
                           15600
  F9 B5 B1051.4
                           15600
  F9 B5 B1049.5
                           15600
  F9 B5 B1060.2
                           15600
  F9 B5 B1058.3
                           15600
  F9 B5 B1051.6
                           15600
  F9 B5 B1060.3
                           15600
  F9 B5 B1049.7
                           15600
```

2015 launch records

2015 Launch records can be found below:

Task 9: List the records which will display the month names, failure landing in drone ship outcomes, booster versions, and launch_sites for the months in 2015

```
%%sql

SELECT MONTHNAME("Date") AS Month, "Booster_Version", "Launch_Site", "Landing_Outcome"
FROM SPACEXTBL
WHERE (("Landing_Outcome" LIKE '%Failure%') AND YEAR("Date")=2015);

* ibm_db_sa://fgl32023:***@dashdb-txn-sbox-yp-dal09-04.services.dal.bluemix.net:50000/BLUDB
Done.

MONTH Booster_Version Launch_Site Landing_Outcome

January F9 v1.1 B1012 CCAFS LC-40 Failure (drone ship)

April F9 v1.1 B1015 CCAFS LC-40 Failure (drone ship)
```

Rank success count between 2010-06-04 and 2017-03-20

The success ranking can be found below:

Task 10: Rank the count of successful landing_outcomes between date 2010-06-04 and 2017-03-20 in descending order

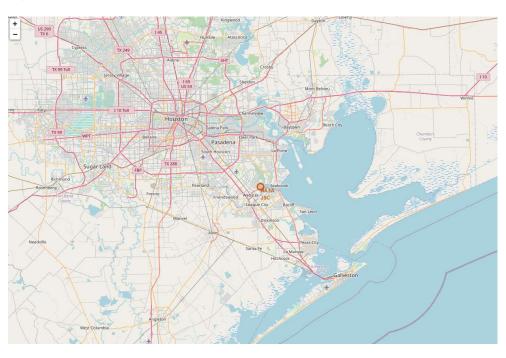
```
SELECT "Landing_Outcome", COUNT(*) AS TOTAL
FROM SPACEXTBL
WHERE (("Landing_Outcome" LIKE '%Success%') AND ("Date" > '2010-06-04') AND ("Date" < '2017-03-20'))
GROUP BY "Landing_Outcome"
ORDER BY "Landing_Outcome" DESC;

* ibm_db_sa://fgl32023:***@dashdb-txn-sbox-yp-dal09-04.services.dal.bluemix.net:50000/BLUDB
Done.

Landing_Outcome total
Success (ground pad) 3
Success (drone ship) 5</pre>
```

NASA Map

Here is the location of NASA in Houston, TX:



SpaceX Launch Sites

This map displays the location of SpaceX launch sites in blue.



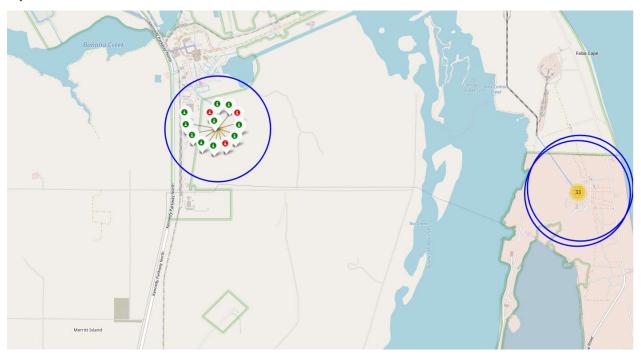
Launch History

This map displays the number of launches at each launch site.



Launch Site Successes and Failures

This map displays the success or failure of each launch at each launch site:



<Dashboard screenshot 3>

Replace < Dashboard screenshot 3> title with an appropriate title

Show screenshots of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider

Explain the important elements and findings on the screenshot

ALGORITHM	JACCARD TRAIN ACCURACY SCORE (.accuracy_score())	JACCARD TEST ACCURACY SCORE (.accuracy_score)	TEST ACCURACY R-SQUARED SCORE (.score())	GRIDSEARCHCH Score (.BEST_SCORE)
Logistic Regression	.875	.8334	.8334	.8464
Support Vector Machine	.889	.8334	.8334	.8484
Decision Tree	.8611	.8333	.8333	.8768
K-Nearest Neighbors	.8611	.8334	.8334	.848

Classification Accuracy

All four models had roughly the same level of accuracy, so any would be a good model to use for prediction purposes.

```
: # plot confusion matrix
  yhat = knn cv.predict(X test)
  plot_confusion_matrix(Y_test,yhat)
                   Confusion Matrix
                                                   - 12
                                                    - 10
  True labels
did not land
                                    12
    landed
            did not land
                                   land
                     Predicted labels
: print("LR jaccard train score: ", metrics.accuracy_score(Y_train,knn_cv.predict(X_train) ))
  print("LR jaccard test score: ", metrics.accuracy_score(Y_test, yhat))
  print("LR R-squared score: ", knn_cv.score(X_test, Y_test))
  print("LR gridsearch score", knn_cv.best_score_)
  LR jaccard train score: 0.8611111111111112
  LR jaccard test score: 0.83333333333333334
  LR R-squared score: 0.83333333333333334
  LR gridsearch score 0.8482142857142858
```

Confusion Matrix

All four models performed roughly the same and had the same confusion matrix. False negatives were the biggest problem with each model.

CONCLUSION



Conclusion:

- The models we used are reliable for predicting future successful launches based off historical data.
- The EDA indicates that certain parameters help make launches more successful, such as:
 - launch site VAFB SLC 4E having a higher success rate than the other launch sites
 - launches with higher flight numbers generally being more successful)
 - launches with ES-L1 orbit type are much more successful than those with GTO orbit types
 - SpaceX is learning from past mistakes, as indicated by the fact that the rate of successful launches has increased over time on average.

APPENDIX

```
print("LR jaccard train score: ", metrics.accuracy_score(Y_train,knn_cv.predict(X_train) )) #
print("LR jaccard test score: ", metrics.accuracy_score(Y_test, yhat))
print("LR R-squared score: ", knn_cv.score(X_test, Y_test))
print("LR gridsearch score", knn_cv.best_score_)
```

[hide] Flight No.	Date and time (UTC)	Version, Booster ^[b]	Launch site	Payload ^[c]	Payload mass	Orbit	Customer	Launch outcome	Booster landing				
78	7 January 2020, 02:19:21 ^[492]	F9 B5 △ B1049.4	CCAFS, SLC-40	Starlink 2 v1.0 (60 satellites)	15,600 kg (34,400 lb) ^[5]	LEO	SpaceX	Success	Success (drone ship)				
	Third large batch and sec	ond operational flight o	of Starlink constella	tion. One of the 60 satellites included a test coating	to make the satellite less reflective, and	thus less likely to inter	fere with ground-based astronomical of	bservations.[493]					
	19 January 2020, 15:30 ^[494]	F9 B5 △ B1046.4	KSC, LC-39A	Crew Dragon in-flight abort test ^[495] (Dragon C205.1)	12,050 kg (26,570 lb)	Sub-orbital ^[496]	NASA (CTS)[497]	Success	No attempt				
79	site. The test was previou	sly slated to be accom	plished with the Cn	e capsule fired its SuperDraco engines, reached an ew Dragon Demo-1 capsule, [498] but that test articl synamic forces after the capsule aborted. [500] First	e exploded during a ground test of Super	Draco engines on 20 A	April 2019. ^[419] The abort test used the	capsule originally					
80	29 January 2020, 14:07 ^[501]	F9 B5 △ B1051.3	CCAFS, SLC-40	Starlink 3 v1.0 (60 satellites)	15,600 kg (34,400 lb) ^[5]	LEO	SpaceX	Success	Success (drone ship)				
	Third operational and four	th large batch of Starli	ink satellites, deploy	yed in a circular 290 km (180 mi) orbit. One of the f	lairing halves was caught, while the other	was fished out of the	ocean.[502]						
81	17 February 2020, 15:05 ^[603]	F9 B5 △ B1056.4	CCAFS, SLC-40	Starlink 4 v1.0 (60 satellites)	15,600 kg (34,400 lb) ^[5]	LEO	SpaceX	Success	Failure (drone ship)				
				a new flight profile which deployed into a 212 km x ata. [505] This was the first time a flight proven boos		instead of launching int	o a circular orbit and firing the second	stage engine twice	. The first stage				
	7 March 2020, 04:50 ^[506]	F9 B5 △ B1059.2	CCAFS, SLC-40	SpaceX CRS-20 (Dragon C112.3 △)	1,977 kg (4,359 lb) ^[507]	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)				
-				n ESA platform for hosting external payloads onto I y part. ^[509] it was SpaceX's 50th successful landing					e failure. SpaceX				
	18 March 2020, 12:16 ^[510]	F9 B5 △ B1048.5	KSC, LC-39A	Starlink 5 v1.0 (60 satellites)	15,600 kg (34,400 lb) ^[5]	LEO	SpaceX	Success	Failure (drone ship)				
	shut down of an engine, to	This operational launch of Starrisk satellities. It was the first strage booster flew for a fifth time and the second time the fairings were reused (Starrink flight in May 2019). Sharrisk satellities in was the first strage burn, the booster suffered premature with down of an engine, the first of a Merin 10 variant and strat since the CRS-1 mission in October 2012. However, the payload still reached the targeted orbit. (Sharrink flight in May 2019). This was the second Starlink launch booster landing failure in a row, later revealed to be caused by resolated desiration fail failured in a sensor. (Sharrink launch booster landing failure in a row, later revealed to be caused by resolated desiration fail failured in a sensor. (Sharrink launch booster landing failure in a row, later revealed to be											
84	22 April 2020, 19:30 ^[514]	F9 B5 △ B1051.4	KSC, LC-39A	Starlink 6 v1.0 (60 satellites)	15,600 kg (34,400 lb) ^[5]	LEO	SpaceX	Success	Success (drone ship)				

